Overview of the NTCIR-10 1CLICK-2 Task

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

Matthew Ekstrand-Abueg
Northeastern University
mattea@ccs.neu.edu

Virgil Pavlu
Northeastern University
vip@ccs.neu.edu

Tetsuya Sakai
Microsoft Research Asia
tetsuyasakai@acm.org

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

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

ABSTRACT

This is an overview of the NTCIR-10 1CLICK-2 task (the second One Click Access task). Given a search query, 1CLICK aims to satisfy the user with a single textual output instead of a ranked list of URLs. Systems are expected to present important pieces of information first and to minimize the amount of text the user has to read. We designed English and Japanese 1CLICK tasks, in which 10 research teams (including two organizers' teams) participated and submitted 59 runs for Main tasks and a Query Classification subtask. We describe the tasks, test collection, and evaluation methods, and then report official results for NTCIR-10 1CLICK-2.

1. INTRODUCTION

In contrast to traditional Web search, which requires the user to scan a ranked list of URLs, visit individual URLs and gather pieces of information he needs, 1CLICK aims to satisfy the user with a single textual output, immediately after the user clicks on the SEARCH button. Participating systems are expected to present important pieces of information first and to minimize the amount of text the user has to read. This type of information access is particularly important for mobile search. The systems are thus expected to search the Web and return a multi-document summary of retrieved relevant Web pages that fits a small screen.

The first 1CLICK task (1CLICK-1) was proposed as a *pilot* task in the previous NTCIR (NTCIR-9) [22], which dealt with Japanese queries only. The second 1CLICK task, 1CLICK-2[1], has expanded its language scope to English and Japanese, and included Main tasks (given a query, return a single textual output) as well as a Query Classification subtask (given a query, return the query type).

1CLICK-2 attracted 10 research teams (including two organizers' teams) from five countries: Japan, China, U.S.A., Canada, and the Netherlands. Tables 1 and 2 provide a list of the NTCIR-10 1CLICK-2 participants with the number of English/Japanese Main task and Query Classification subtask submissions. The total number of English/Japanese Main task submissions is 38, which is almost four times as many as that of 1CLICK-1.

Table 3 shows important dates of NTCIR-10 1CLICK-2. We first released sample queries and nuggets (important pieces of information for each query) to help people better understand the 1CLICK-2 task. We then released test queries and a document collection from which participants are expected to generate a single textual output. We received runs from participants on October 31 in 2012. Although the original result release was scheduled on February 28, we could release only Japanese evaluation results in time, and released English evaluation results on April 7, 2013. We apologize for the late English evaluation result release.

Table 3: Important dates of NTCIR-10 1CLICK-2.

date	
April 30, 2012	Sample queries and nuggets released
Aug 31, 2012	Test queries and a document collection released
Oct 31, 2012	Run submissions due
Nov 2012 - Jan 2013	Evaluation
Feb 01, 2013	Very early draft overview released
Feb 28, 2013	Japanese evaluation results released
Apr 1, 2013	Almost-camera-ready paper released
Apr 7, 2013	English evaluation results released

The remainder of this paper is structured as follows. Section 2 describes the details of the Main tasks and Query Classification subtask. Section 3 introduces a test collection that constitutes queries, nuggets, *iUnits*, *vital strings*, and a document collection. Section 4 explains semi-automated methods of extracting nuggets and vital strings for English subtasks. Section 5 describes the evaluation methods we used. Sections 6 and 7 report on our official evaluation results of the Main tasks and Query Classification subtask, respectively. Finally, Section 8 concludes this paper.

[1] 1CLICK-2 homepage: `http://research.microsoft.com/1CLICK/`

Table 1: NTCIR-10 1CLICK-2 participants with the number of English Main task and Query Classification subtask submissions.

team name	Main tasks	Query Classification subtask	organization
KUIDL	4	2	Kyoto University, Japan
NSTDB	6	0	Nara Institute of Science and Technology, Japan
NUIR	8	0	Northeastern University, U.S.A.
NUTKS	0	6	Nagaoka University of Technology, Japan
udem	4	0	University of Montreal, Canada
ut	2	2	University of Twente, the Netherlands
Total	24	10	

Table 2: NTCIR-10 1CLICK-2 participants with the number of Japanese Main task and Query Classification subtask submissions.

team name	Main tasks	Query Classification subtask	organization
HUKB	0	2	Hokkaido University, Japan
KUIDL	4	2	Kyoto University, Japan
MSRA	4	1	Microsoft Research Asia, P.R.C.
NUTKS	0	6	Nagaoka University of Technology, Japan
ORG	3	0	Organizers' team
TTOKU	3	0	Tokyo Institute of Technology, Japan
Total	14	11	

2. TASKS

1CLICK-2 comprises the Main tasks and Query Classification subtask. This section explains the two types of tasks, and the format of input and output.

2.1 Main Tasks

The Main tasks are similar to the 1CLICK-1 task, where systems are required to return a single text output for a given query. Participants can choose *device* and *source* types that suit their systems for each Main task run.

2.1.1 Input

The input for the Main tasks is a query file in which each line is of the following form:

<queryID>[TAB]<querystring>

In the 1CLICK-1 task, we selected four types of queries (CELEBRITY, LOCAL, DEFINITION, and QA) based on findings from a mobile query log study [13]. We used more fine-grained query types in this round, and for each type we assume that the user has the following information needs:

ARTIST (10) user wants important facts about musicians, novelists etc. who produce works of art;

ACTOR (10) user wants important facts about actors, actresses, TV personalities etc.;

POLITICIAN (10) user wants important facts about politicians;

ATHLETE (10) user wants important facts about athletes;

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

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 an unfamiliar term, an idiom etc.;

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

The number of queries for each query type is shown in parentheses. Thus, we used a total of 100 queries for the English subtask, and another 100 for the Japanese subtask.

2.1.2 Output

For each input query, 1CLICK systems are expected to produce a plain text output of X characters, excluding punctuation marks, special symbols etc. We call this the X-string. The length of X-strings is limited as follows:

- For English, 1,000 characters for DESKTOP run and 280 characters for MOBILE run; and
- For Japanese, 500 characters for DESKTOP run and 140 characters for MOBILE run.

The DESKTOP run roughly corresponds to five search engine snippets that the user can typically view without scrolling the browser window (i.e., those "above the fold."), while the MOBILE run approximates a mobile phone display size.

The longer length limit for English subtasks reflects the fact that Japanese characters hold more information than English ones. The average text reading speed in English is approximately 250-300 words per minute and words are approximately 4 characters long on average, while the average text reading speed of a Japanese person is approximately 400-600 words.

In addition to the *device* type (DESKTOP or MOBILE), participants can select *source* type for each run. There are three source types:

Mandatory (MAND)
organizers provided Web search results and their page contents for each query. Participants may use only this information to generate X-strings.

Oracle (ORCL)
organizers provided a list of relevant pages for each query[2], which are a subset of the pages provided for Mandatory runs. Participants can use the data either wholly or partially to generate X-strings. If this data set is used in any way at all, the run is considered an Oracle run.

Open (OPEN)
participants may choose to search the live Web on their own to generate X-strings. Any run that does not use the oracle data but uses at least some privately-obtained Web search results is considered an Open run, even if it also uses the pages provided for Mandatory runs.

Note that participants were required to generate X-strings from the content in the given resources, but we did not restrict anything except the source of X-strings. Participants can use external knowledge sources (e.g. WordNet) or publicly available resources (e.g. Wikipedia) to summarize given page contents. However, we highly recommended that participants did not use unreproducible or proprietary resources such as the live Web and query logs in Mandatory and Oracle runs.

Participants were asked to submit at least one Mandatory run for the Main task to enhance the repeatability and comparability of 1CLICK experiments. Oracle runs were designed for participants who are interested mainly in search result summarization, and

[2]Relevant pages are pages where organizers find at least one nugget

Open runs were designed for ones who want to use unreproducible but useful data for X-string generation (e.g. the live Web).

As there are two language, two device, and three source types, twelve types of run were possible. Participants are required to specify the run type by the run file name and to follow the following naming convention:

<team>-<lang>-<device>-<source>-<priority>.tsv

where <team> is the team name, <lang> is either E (English) or J (Japanese), <device> is either D (DESKTOP) or M (MOBILE), and <source> is either MAND (Mandatory), ORCL (Oracle), or OPEN (Open). <priority> is a unique integer for each team's run starting from 1, which represents the evaluation priority of a run file. Some example run names for a team "MSRA" would be:

- MSRA-E-D-MAND-1.tsv (English DESKTOP Mandatory run)
- MSRA-E-M-OPEN-2.tsv (English MOBILE Open run)
- MSRA-J-D-ORCL-3.tsv (Japanese DESKTOP Oracle run)

Each team can submit up to four runs for Japanese and six runs for English.

Each run file begins with exactly one system description line, which should be in the following format:

SYSDESC[TAB]<one-sentence system description>

Below the system description line, there must be an output line for each query. Each output line should contain an X-string. The required format is:

<queryID>[TAB]OUT[TAB]<X-string>

Each output line should be followed by at least one SOURCE line. These lines represent a document from which the X-string is generated. Each SOURCE line must be in the following format:

<queryID>[TAB]SOURCE[TAB]<source>

where <source> is either a URL or the filename of a page provided for Mandatory runs. These lines were used for investigating what kinds of knowledge sources the participating teams utilized.

An example of the content of a run file is as follows:

SYSDESC[TAB]baseline 1CLICK system
1C2-E-0001[TAB]OUT[TAB]On June 25, 2009 A...
1C2-E-0001[TAB]SOURCE[TAB]http://.../
1C2-E-0001[TAB]SOURCE[TAB]http://.../
...

2.2 Query Classification Subtask

As the Main tasks demand much effort for new participants, we also devised a relatively easy but important subtask for 1CLICK systems, namely, Query Classification subtask. The Query Classification subtask requires systems to predict the query type for a given query, i.e. multiclass query classification into ARTIST, ACTOR, POLITICIAN, ATHLETE, FACILITY, GEO, DEFINITION, and QA. Main task participants whose systems involve query classification are encouraged to participate in this subtask to *componentize* evaluation.

The input of the Query Classification subtask is the same as that of the Main tasks: a query file, which contains pairs of a query ID and a query string as explained earlier.

The file name of each run in the Query Classification subtask is of the following format[3]:

<team>-QC-<priority>.tsv

where <team> is the team name, and <priority> is a unique integer for each team's run starting from 1. Participants can submit as many Query Classification runs as they like.

Each line in the run file contains the following two fields:

<queryID>[TAB]<querytype>

where <querytype> is a query type predicted by the system, which must be one of the following eight types: ARTIST, ACTOR, POLITICIAN, ATHLETE, FACILITY, GEO, DEFINITION, and QA. Note that there are no source types such as Mandatory, Oracle, and Open in the Query Classification subtask[4].

[3]It should have been better to specify the language (English or Japanese) by the file name, even though organizers can guess the language by looking at the query ID.

[4]It also should have been better to distinguish systems that use proprietary data and the others in the Query Classification subtask.

3. TEST COLLECTION

The NTCIR-10 1CLICK-2 test collection includes queries, nuggets, and a document collection for Mandatory runs as well as a list of relevant documents for Oracle runs. Nuggets are relevant pieces of information, which are prepared for each query and used for evaluating X-strings. The more important nuggets appear in earlier parts of the X-string, the higher the X-string is evaluated. To judge which nuggets match an X-string, we break a nugget into smaller units, iUnits and vital strings, and used an iUnit as a unit for matching in Japanese subtasks and a vital string for matching in English subtasks. The NTCIR-10 1CLICK-2 test collection thus consists of the queries, nuggets, and document collection as well as iUnits and vital strings. We describe the details of those components in the following sections.

3.1 Queries

The NTCIR-10 1CLICK-2 test collection includes 100 English and 100 Japanese queries (see Appendix A for the complete lists). English queries were selected from Bing mobile search query logs[5], of which 15 overlap with Japanese queries, allowing for cross-language comparison. The overlap queries consist of one query from ARTIST and ACTOR, two queries from POLITICIAN and ATHLETE, and three queries from FACILITY, DEFINITION, and QA.

In order to make the task more interesting, and to discourage simply returning the first paragraph of a Wikipedia entry for the given entity or the snippets returned by the search engine, many of the English queries were given a specifier as well, e.g. JFK conspiracy theory and Kofi Annan Syria. This, along with differing popularity of entities, contributed to the varying level of difficulty of the queries. The specified queries are ones marked as "SPEC" in the query list in Appendix A.

Japanese queries were derived from English queries, NTCIR-10 INTENT-2 Japanese queries, and Yahoo! Chiebukuro[6]. See Appendix A for the details of query sources.

3.2 Nuggets

The 1CLICK -2 task focuses on evaluating textual output based on nuggets rather than document relevance. Nuggets have been used for evaluating complex question-answering [8, 15] and document retrieval [18], and were recently used for 1CLICK task evaluation [21]. The nugget in 1CLICK-2 is defined as a sentence relevant to the information need for a query, and was used to evaluate the quality of X-strings by identifying which nuggets are present in the content of X-strings. Table 4 shows examples of the nugget, where each nugget is represented as a tuple of a nugget ID, a nugget, and an URL from which the nugget was extracted.

For English subtasks, organizers extracted nuggets from relevant documents by means of semi-automatic extraction. The details of the semi-automatic extraction are explained in Section 4.

For Japanese subtasks, native Japanese speakers in the organizer team identified relevant document from a document collection for Mandatory runs, and manually extracted relevant sentences as nuggets. We extracted 3,927 nuggets for 100 Japanese queries (39.2 nuggets per query).

3.3 iUnits

Nuggets vary considerably in length, and can include dependent multiple pieces of information with different importance. Due to those natures of nuggets, we found it difficult to determine whether a nugget match an X-string or not in 1CLICK-1. Thus, we break nuggets into *relevant*, *atomic*, and *dependent* pieces of information and used them as a unit for text matching. Those pieces of information extracted from nuggets are what we call *iUnits*. iUnits were used for evaluating Japanese runs, where assessors manually checked for the presence and recorded the position of each iUnit in the X-string.

We describe the three properties of iUnits below, i.e. *relevant*, *atomic*, and *dependent*. *Relevant* means that an iUnit provides useful factual information to the user on its own. Thus, it does not require other iUnits to be present in order to provide useful information. For example:

(1) Tetsuya Sakai was born in 1988.

(2) Tetsuya Sakai was born.

If the information need is "Who is Tetsuya Sakai?", (2) alone is probably not useful and therefore this is not an iUnit.

Atomic means that a Japanese iUnit cannot be broken down into multiple iUnits without loss of the original semantics. Thus, if it is broken down into several statements, at least one of them does not pass the relevance test. For example:

(1) Takehiro Yamamoto received a PhD from Kyoto University in 2011.

(2) Takehiro Yamamoto received a PhD in 2011.

(3) Takehiro Yamamoto received a PhD from Kyoto University.

(4) Takehiro Yamamoto received a PhD.

(1) can be broken down into (2) and (3), and both (2) and (3) are *relevant* to the information need "Who is Takehiro Yamamoto?". Thus, (1) cannot be an iUnit, but (2) and (3) are iUnits. (2) can be further broken down into (4) and "Takehiro Yamamoto received something in 2011". However, the latter does not convey useful information for the information need. The same goes for (3). Therefore, (2) and (3) are valid iUnits and (4) is also an iUnit.

Dependent means that a Japanese iUnit can entail other iUnits. For example:

(1) Takehiro Yamamoto received a PhD in 2011.

(2) Takehiro Yamamoto received a PhD.

(1) entails (2) and they are both iUnits.

Organizers manually extracted iUnits from nuggets. The total number of iUnits is 7,672 (76.7 iUnits per query). Thus, 1.95 iUnits were extracted from a nugget. A set of iUnits for query 1C2-J-0001 "倉木麻衣" is shown in Table 5. The column "entail" indicates a list of iUnits that are entailed by the iUnit. For example, iUnit I014 entails I013, and iUnit I085 entails iUnits I023 and I033. The entailment was manually judged by organizers. A *semantics* is the factual statement that the iUnit conveys. This is used by assessors to determine whether an iUnit is present in the X-string or not. The *vital string* is explained in the next subsection.

Having extracted iUnits from nuggets, four of the organizers gave the importance to each iUnit on five point scale (very low (1), low (2), medium (3), high (4), and very high (5)). iUnits were randomly ordered and their entailment relationship was hidden during the voting process. After the voting, we revised iUnit's importance so that iUnit i entailing iUnits $e(i)$ receives the importance of only i excluding that of $e(i)$. This revision is necessary because the presence of iUnit i in an X-string entails that of iUnits $e(i)$, resulting

[5] http://www.bing.com/

[6] Japanese Yahoo! Answer. http://chiebukuro.yahoo.co.jp/

Table 4: Nuggets for Japanese query 1C2-J-0001 “倉木麻衣”.

nuggetID	nugget	URL
S005	99.10 月、16 歳で “ Mai・k ” 名義の『Baby I Like』で全米デビュー。同年 12 月 8 日『Love, Day After Tomorrow』で倉木麻衣、日本デビュー。	http://example.com/
S008	血液型 B 型	http://example.com/
S012	職業 歌手	http://example.com/
S022	2005 年 立命館大学を卒業	http://example.com/
S023	第 15 回日本ゴールドディスク大賞で『delicious way』がロック・アルバム・オブ・ザ・イヤーを、「Secret of my heart」がソング・オブ・ザ・イヤーを受賞。	http://example.jp/

Table 5: iUnits for Japanese query 1C2-J-0001 “倉木麻衣”.

iUnitID	entail	nuggetID	semantics	vital string
I011		S005	1999 年日本デビュー	1999 年日本デビュー
I012		S008	血液型 B 型	血液型 B 型
I013		S022	立命館大卒	立命館大卒
I014	I013	S022	2005 年立命館大卒	2005 年
I017		S012	職業 歌手	歌手
I023		S023	第 15 回日本ゴールドディスク大賞ソング・オブ・ザ・イヤー受賞	第 15 回日本ゴールドディスク大賞ソング・オブ・ザ・イヤー受賞
I033		S023	シングル Secret of my heart	シングル Secret of my heart
I085	I023, I033	S023	第 15 回日本ゴールドディスク大賞で「Secret of my heart」がソング・オブ・ザ・イヤーを受賞	

in duplicative counting of the importance of $e(i)$ when we take into account the importance of both i and $e(i)$.

For example, suppose that there are only four iUnits:

(1) Ichiro was a batting champion (3).

(2) Ichiro was a stolen base champion (3).

(3) Ichiro was a batting and stolen base champion (7).

(4) Ichiro was the first player to be a batting and stolen base champion since Jackie Robinson in 1949 (8).

where (4) entails (3), and (3) entails both (1) and (2). A parenthesized value indicates the importance of each iUnit. Suppose that an X-string contains (4). In this case, the X-string also contains (1), (2), and (3) by definition. If we just sum up the weight of iUnits in the X-string, the result is 21 (= 3 + 3 + 7 + 8), where the importance of (1) and (2) is counted three times and that of (3) is counted twice. Therefore, it is necessary to subtract the importance of entailing iUnits to avoid the duplication; in this example, thus, the importance of iUnits becomes 3, 3, 4 (= 7 - 3), and 1 (= 8 - 7), respectively.

More formally, we used the following equation for revising the importance of iUnit i:

$$w(i) - \max(\{w(j)|j \in e(i)\}), \quad (1)$$

where $w(i)$ is the sum of the importance of iUnit i given by four organizers. Note that iUnits $e(i)$ in the equation above are ones entailed by iUnit i and the entailment is *transitive*, i.e. if i entails j and j entails k, then i entails k.

In the revision process, some iUnits received the importance of zero or less. There are two possible explanations for negative scores:

(1) The iUnit was underestimated in the voting, or

(2) The iUnit is not important enough to be an iUnit.

In 1CLICK-2, we assumed the second reason, and removed iUnits whose importance was zero or less as those iUnits do not satisfy the *relevant* property.

3.4 Vital Strings

A vital string is a minimally adequate natural language expression and extracted from either nuggets or iUnits. For Japanese subtasks, a vital string was extracted from a iUnit. This approximates the minimal string length required so that the user who issued a particular query can read and understand the conveyed information. The vital string of iUnit i that entails iUnits $e(i)$ does not include that of iUnits $e(i)$ to avoid duplication of vital strings, since if iUnit i is present in the X-string, iUnits $e(i)$ are also present by definition. For example, the vital string of iUnit I014 does not include that of iUnit I013 as shown in Table 5. Even the vital string of I085 is empty as it entails iUnits I023 and I033.

For English subtasks, vital strings were derived directly from nuggets. Unlike the use of vital strings for minimal length approximation in Japanese subtasks, vital strings in the English subtasks were used as the unit for automatic matching with the X-strings. The vital string extraction from nuggets and their matching against X-strings are described in Section 4.

3.5 Documents

To provide participants with a document collection for Mandatory runs, we downloaded 500 top-ranked documents that were returned by Bing search engine in response to each query. This download was conducted on July 4, 2012. The title, summary provided by the search engine, URL, and rank of documents were recorded and released along with the document collection. As we downloaded documents cached by the search engine and failed to access some of the caches, the number of downloaded documents per query is fewer than 500. The average number of documents for English queries is 269.8 and that for Japanese queries is 390.0.

We extracted nuggets, iUnits, and vital strings from the document collection for Mandatory runs, and created a list of relevant

documents, which are defined as documents from which at least one nugget was extracted. The list was released for Oracle runs, where participants can use the relevant document list to generate X-strings.

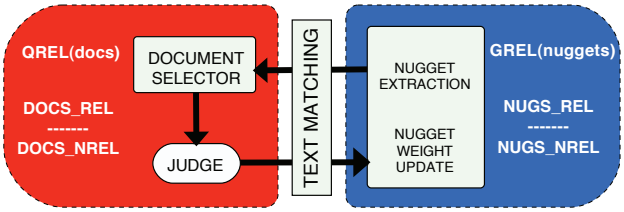


Figure 1: The overall design of the English subtask assessment process: Iteratively, documents are selected and assessed, and nuggets are extracted and [re]weighted.

4. NUGGET AND VITAL STRING EXTRACTION FOR ENGLISH SUBTASKS

While the Japanese 1CLICK-2 focuses on the iUnit as the evaluation unit, the English subtasks use functionally equivalent, although slightly different looking text units. The reason for this is primarily that an iUnit is matched manually by an assessor to a summary. One goal of the English subtasks is to develop semi-automated methods for performing the extraction of gold text units and for matching them with the participant summaries. As such, we begin, as the Japanese subtasks do, by extracting source sentences, or nuggets, from relevant documents. These are then broken up into vital strings, which are, like iUnits, small, relevant textual units. However the goal of a vital string is to be as close to atomic as possible, in that a match to a vital string should indicate the presence of one piece of information. An iUnit may represent a single primary piece of information along with peripheral, but relevant, details. We represent these details as vital string dependencies.

Take for example a nugget from the query "Michael Jackson Death": "Murray tried to revive Jackson for five to ten minutes, at which point he realized he needed to call for help." In this, we have the primary piece of information in the sentence, "Murray tried to revive Jackson." This would be the primary vital string. Next we have the duration, "for five to ten minutes," as a secondary vital string dependent on the existence of the first. Additionally we have "realized he needed to call for help" as an additional secondary vital string. For matching sake, we also remove stop words and query terms, as they do not add useful information to the vital string, except when being viewed by humans.

4.1 Framework

We are using a framework of mutual, iterative reinforcement between nuggets and documents, most similar to Active Learning [3, 24] mechanisms. The human feedback (relevance assessment) is given *iteratively* on the documents/data. Thus, while the assessor judges the document (as it is standard IR procedure), our "back-end" system infers the good nuggets.

Figure 1 illustrates the framework, with the MATCHING being the reinforcing procedure from documents to nuggets and vice-versa:

- the iterative loop selects documents based on the current notion of relevance (as expressed by current set of nuggets G); documents are judged and put into the set of documents (docs QREL);
- for relevant documents, all nuggets are extracted and added to G
- all nugget weights/qualities are updated given the relevant and non-relevant documents judged so far and how well they match each nugget

This framework uses four components that can be designed somewhat independently.

We present our chosen implementations for SELECTION, MATCHING, EXTRACTION, and DISTILLATION, noting that for each of these we settled on these methods after trying various ideas. While ours work well, each can be independently replaced by more suited techniques for specific tasks. A fifth component, not explored in this paper, is the JUDGE model. Here we assume the judge follows the traditional NIST assessor, simply labeling binary documents as relevant (contains something of some relevance) or non-relevant; a more sophisticated judge can review and modify the nuggets extracted, categorize documents, use graded relevance, annotate important keywords, or classify the query.

4.2 Syntactic Text Matching

The text MATCHING between nuggets and documents is both the syntactic problem of measuring whether the nugget and the document describe the same people, places, locations, dates, facts, actions, etc and the semantic, or textual entailment problem of reflecting the fact that the document entails (presents the semantical information found in) the nugget.

As presented in [18], the matching algorithm is based on a variant of *shingle matching*, which is often used in near-duplicate detection [4, 5]. A shingle is a sequence of k consecutive words in a piece of text. For example, after stopwording, the nugget `"John Kennedy was elected president in 1960"` has the following shingles for $k = 3$: (`John Kennedy elected`), (`Kennedy elected president`), and (`elected president 1960`).

In contrast to standard *shingle matching* used for duplicate detection, we do not require all shingle words to be present in the matching document in the same order or contiguously. Our method is inspired by near-duplicate detection, but is in fact quite different. High scores indicate a match of known relevant information, not necessarily of redundant or duplicate text. Furthermore, while a known nugget is required to be present in a document for a good match, the document often contains new/unknown relevant information as well. For more information on this procedure, see [19].

Nugget matching: To obtain a matching score for nugget n against document d, we average the scores for each of nugget shingles:

$$\begin{aligned} nuggetmatch &= M(n,d) = \\ &= \frac{1}{\#shingles} \sum_{s \in shingles(n)} shingleMatch(s,d) \end{aligned}$$

While looking for a good but simple text matching strategy, we considered two bag-of-words possibilities: BM25 and Jaccard's Coefficient; both significantly underperformed the presented shingle technique [19].

4.3 Document Selection

The document SELECTION process decides what documents to be judged next, given the current nuggets and their weights, matches, and possibly other contextual information such as the position of documents in retrieved lists, specifics of the task, or desired balance between exploration and exploitation.

At each iteration of the document - nugget loop, one or more documents are selected and immediately assessed by the judge. The selection has two parts: First, a ranking function ranks all candidate documents based on a score obtained by matching existing nuggets: at round r, G is the current set of nuggets $n \in G$, each with current quality weight q_n. Then the document score is a dot product of

quality and matching

$$DocScore(d) = \sum_{n \in G} q_n * M(n, d)$$

Since some documents can equally match the nugget set (a zero match in the worst case), we add to the document score a retrieval value considering the positions of the documents in various retrieved lists (submitted IR systems in TREC case). Such methods have been developed before [2, 6, 7]; we use the method described in [2]. Second, a probabilistic procedure picks a number of documents from the ranking. We found that a good balance between exploitation and exploration can be achieved by sampling from a scaled geometric distribution

$$\alpha(K)p(1-p)^{r-1}$$

associated with the document ranks r (not scores), with the geometric base $p = 0.4$ found empirically.

4.4 Nugget Extraction and Weighting.

Finally, the nugget EXTRACTION is responsible for deciding what nuggets to consider from judged-relevant documents, and also responsible in deciding how much each nugget currently matters (updating nugget weight).

Nuggets are extracted *only from relevant documents* by a simple NLP script which looks for sentences; stemming and stopwording is performed on the nuggets to improve matching specificity. We treat the nuggets as *experts* and at each iteration we compute the *current nugget quality* q_n, an *importance weight* used by the document selector when computing document scores. These weights are only an indication of the current belief that a nugget is relevant: if matching documents are proven non-relevant, the weight becomes insignificant; a nugget only maintains a high weight across many iterations if it consistently matches relevant documents.

The weighting mechanism is adapted from experts combination/online learning, where the problem has been exhaustively studied. We use the "Hedge" algorithm [10], the internal engine of AdaBoost and Rankboost (but without boosting), which has been shown to have good theoretical properties [23] and fast implementations. A relevant document d increases the weight of the matching nugget n based on the matching score:

$$q_n^{new} = q_n^{old} / \beta_1^{M(n,d)}, normalized$$

while non-relevant document decreases the weight:

$$q_n^{new} = q_n^{old} * \beta_2^{M(n,d)}, normalized$$

Some nuggets are extracted very early in the loop, while others enter G very late; for consistency, the initial weight of each nugget is computed as if the nugget existed in G all along. The two beta values are found by trial and error and for our systems are set at $\beta_1 = 0.8$ and $\beta_2 = 0.5$. This hedge variant with β values associated with each feedback outcome (relevant, non-relevant) was first described in [1]. Other minor heuristics are used to prevent nugget weights from being extremely large. A particularly important feature of Hedge is its adaptability [10]: nugget quality is high as long the nugget brings in relevant documents; after that, it decreases if non-relevant documents match the nugget, or the nugget becomes irrelevant (irrespective of quality q_n) if no other documents match it.

4.5 Vital Strings

The vital strings are the result of a DISTILLATION from nuggets using Stanfords' NLP pipeline and some heuristics. A massive manual editing effort is necessary in order obtain fact-based vital strings breakups of each iUnit. In the following example, we see an iUnit followed by the extracted vitals strings:

```
   1C2-E-TEST-0001-1.10 "[10] Murray tried
to revive Jackson for five to ten minutes,
at which point he
realized he needed to call for help."
1C2-E-TEST-0001-1.10.001 murray tried revive
1C2-E-TEST-0001-1.10.002 five ten minutes
DEP=1C2-E-TEST-0001-1.10.001
1C2-E-TEST-0001-1.10.003 realized he needed
to call help DEP=1C2-E-TEST-0001-1.10.001
```

The first line contains the initial nugget, and the following three lines contain a primary vital string, along with two dependent vital strings. This system allows for construction of concise text for matching, while retaining some important linguistic features in terms of primary facts and peripheral details. Note that these vital strings have many stopwords and query terms removed as these do not add additional intrinsic value to an information unit for a given query.

Once the distillation process is complete, the resulting vital strings can be used to find relevant portions of text within the participant summaries using completely manual, semi-supervised, or fully automatic methods.

5. EVALUATION METHODS

This section explains methods of evaluating X-strings based on iUnits and vital strings. We first identify the presence and position of iUnits or vital strings in the X-string. In parallel, we refine our nugget set, adding ones that appear in submitted X-strings but are not included in the existing nugget set. We finally compute several metrics such as weighted recall, S-measure, T-measure, and S#-measure, based on the match position.

5.1 Matching

In English subtasks, we developed an semi-automatic vital string matching system, and ran it to identify the position of iUnits. In Japanese subtasks, by contrast, we conducted manual iUnit matching where subjects were asked to manually identify the position of iUnits in the X-string.

English subtasks.

In order to achieve high accuracy of the matches between our vital strings and participant summaries, we used a manual matching system with an automatic prior simply to reduce the workload for the assessors. The matching system consisted of three primary tasks: finding matches between existing vital strings and each summary, finding any new vital strings within each summary, and rating the importance of each vital string. These principles are the same as those used in the Japanese subtasks, although the interface and simultaneous method for collecting this data differs from the Japanese system.

In Figure 2 you can see an example of the entailment interface. The left side of the interface contains the summary currently being evaluated, highlighted to show the current matches. The right side contains a scrollable and searchable list of all vital strings for the query, including some context of the vital string if deemed necessary along with its current match positions and judged importance, sorted for important and current match status (high importance first, and matching vital strings first).

In the summary, the blue highlighted text represents automatic matches of vital strings to the summary. These are verified by the assessor during the assessment process. The green matches indicate manually assigned matches. For ease of correlation, hovering over a vital string highlights in yellow its current match in the summary, if one exists. As mentioned, new vital strings can be added as well to ensure full coverage of the information in the summaries.

Although we performed matching on the original summary X-strings regardless of their length, we truncated them to the character limits for the purpose of evaluation, and ignored all matches that spanned beyond the truncation. As the character limit in the guidelines did not include special characters, this truncation and in fact all character position matches were scaled by removing all non-word, non-space characters, and merging sequences of whitespace into a single space.

Japanese subtasks.

In Japanese subtasks, we hired seven assessors who are fairly skilled in Japanese for the iUnit matching task, and asked them to identify the position of each iUnit in the X-string by using the *1CLICKEVAL* system shown in Figure 3. Assessors can select text within the X-string at the left pane for an iUnit listed at the left pane, and click on the "Save" button that pops up under the mouse cursor to record the position of the iUnit. In Figure 3, "is a cartoonist" within the X-string is selected for iUnit "Catoonist" at the top of the right pane. The position of matched text is recorded in the form of "[<starting position>, <ending position>]", e.g. [30, 36] under the first iUnit at the right pane.

Two assessors evaluated each X-string independently in 1CLICK-2 Japanese subtasks, where the order of X-strings was randomized to reduce the order bias. The inter-assessor agreement was considered *good*: 0.747 in terms of average Cohen's Kappa coefficient, where we used the agreement on the presence and absence of an iUnit. The Cohen's Kappa coefficient varied between 0.688 and 0.818, which are almost the same as 1CLICK-1 (from 6.88 to 0.88). We also measured the minimum distance between the ending position of iUnits reported by two assessors, where excluded were iUnits that were not judged as present by one of the two assessors. The mean absolute error and square error were 6.41 and 1280, respectively. The detail of inter-assessor agreement is described in Appendix B. The average time required for an assessor to evaluate an X-string was 131 seconds, which is almost the same as the average assessment time of 151 seconds in 1CLICK-1.

The 1CLICKEVAL system also provides radio buttons for evaluating the readability and trustworthiness of each X-string, as shown at the top of Figure 3. The assessors were asked to evaluate those two criteria on a four point scale: very low (-2), low (-1), high ($+1$), and very high ($+2$). The definitions of the readability and trustworthiness are the same as 1CLICK-1:

Readability.

is to do with coherence and cohesiveness, and how easy it is for the user to read and understand the text. For example, garbled text and the lack of spaces between two unrelated contexts can hurt readability.

Trustworthiness.

means whether the user is likely to believe what it says in the X-string, as well as whether the user is likely to be misled. For example, if the X-string looks as if it was extracted from a source that is clearly not authoritative, it is not trustworthy. Moreover, if what is implied in the X-string is contrary to facts (which can happen, for example, when pieces of information from multiple sources are mixed together), it is not trustworthy.

5.2 Nugget Refinement

We prepared nuggets to cover all the relevant pieces of information before conducting the match process. However, the nugget set does not always suffice to evaluate submitted runs as we might miss some nuggets. Thus, we conducted a nugget refinement process based on submitted runs.

In English subtasks, new vital strings found within submitted X-strings were added to the vital string pool after simple duplicate detection to weed out repeat vital strings.

In Japanese subtasks, we asked the seven assessors to find relevant sentences that are not covered by the prepared iUnits when they were working on the iUnit match evaluation. The total number of nuggets found in the nugget refinement process was 1,002, from which organizers manually extracted 1,501 iUnits. The total number of iUnits is thus 9,173 (7,672 iUnits prepared before the iUnit match process + 1,501 iUnits extracted during the iUnit match process).

5.3 Evaluation Metrics

Several evaluation metrics were calculated based on the presence and position of iUnits or vital strings in the X-string. Since the metrics we used are common between English and Japanese subtasks, we explain the metrics by using only iUnits below.

5.3.1 Weighted recall

Let I be a set of iUnits constructed for a particular query, and

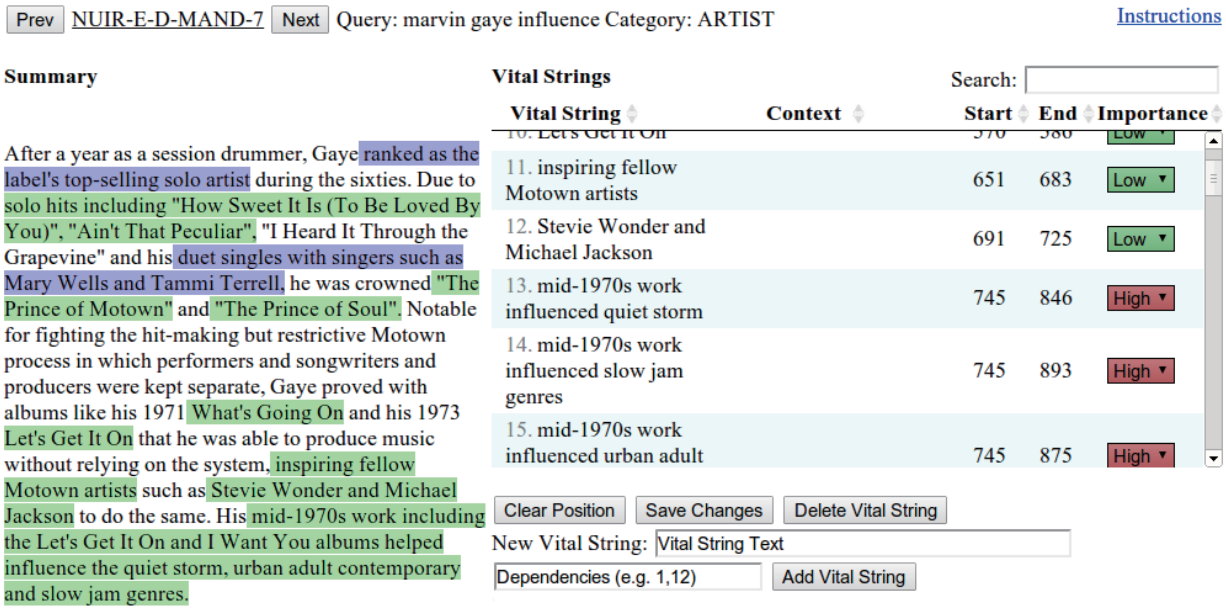


Figure 2: Example Entailment Interface for query *"marvin gaye influence"* in English subtasks.

Table 6: Toy iUnit examples.

ID	vital string	weight	vital string length
I001	gold medalist	2	12
I002	the first gold medalist	3	20

M ($\subset I$) be a set of matched iUnits obtained by comparing the X-string with the prepared iUnits. Weighted recall is defined as follows:

$$\text{Weightedrecall} = \frac{\sum_{i \in M} w(i)}{\sum_{i \in I} w(i)}, \tag{2}$$

where $w(i)$ (or $w(m)$) is the weight of iUnit i (or m).

5.3.2 S-measure

S-measure proposed by Sakai, Kato, and Song was the primary evaluation metric used at 1CLICK-1 [21]. Letting $v(i)$ be the vital string for iUnit $i \in I$, the *Pseudo Minimal Output* (PMO), which is an ideal X-string artificially created for estimating the upper bound per query, was defined by sorting all vital strings by $w(i)$ (first key) and $|v(i)|$ (second key) in the original S-measure definition. In 1CLICK-2, however, we cannot create the PMO in the same way due to the entailment constraint between iUnits. PMO created by sorting vital strings by their weight and length can contain iUnit i without containing iUnits $e(i)$. This PMO violates the entailment constraint as entailed iUnits $e(i)$ must be present within an X-string containing i. Table 6 shows toy examples to explain this problem, where I002 entails I001. I002 is selected as the first part of the PMO since it has the higher weight. If we cannot include I001 in the PMO due to the length limit, the PMO is supposed be invalid because if the PMO containing I002 must include I001 as well if I002 entails I001.

To cope with this problem, we took the following approaches to create the PMO. We first generated *extended* iUnits by merging an iUnit i with entailed iUnits $e(i)$, where the weight and length of extended iUnit i' are defined as the sum of the weight and the sum of the length of iUnits i and $e(i)$, respectively. Thus,

$$w(i') = w(i) + \sum_{j \in e(i)} w(j), \tag{3}$$

$$|v(i')| = |v(i)| + \sum_{j \in e(i)} |v(j)|. \tag{4}$$

As an extended iUnit i' includes iUnits i and $e(i)$, creating PMO by using extended iUnits does not yield an invalid X-string where entailed iUnits are missing despite the presence of an entailing iUnit. For example, an extended iUnit of I002 in Table 6 has $w(i') = 2 + 3 = 5$ and $|v(i')| = 12 + 20 = 32$.

However, another problem arises: an extended iUnit generated from an iUnit entailing many iUnits are likely to receive high weight but likely to be long. PMO created in the same way as 1CLICK-1 includes such *ineffective* iUnits at the beginning of the PMO, and results in underestimation of an ideal X-string. Therefore, we directly maximize the the denominator of S-measure by iteratively adding an extended iUnit that maximizes a term of the denominator shown below:

$$w(i') \max(0, L - \text{offset}^*(v(i'))), \tag{5}$$

where L is a parameter that represents how the user's patience runs out and $\text{offset}^*(v(i))$ is the vital string offset position of iUnit i within the X-string. The algorithm of creating PMO in 1CLICK-2 is as follows:

(1) prepare an empty set O,

(2) letting $\text{offset}(i')$ be $\sum_{i \in O} |v(i)| + |v(i')|$, add an extended iUnit that maximize Equation 5 and remove it from the iUnit set,

(3) repeat (2) while $\sum_{i \in O} |v(i)|$ does not exceed the length limit, and

Figure 3: 1CLICKEVAL system used for iUnit match evaluation in Japanese subtasks.

(4) output O as PMO.

Although the crude assumptions cannot always yield the optimal and coherent X-string as the definition of PMO in 1CLICK-1 cannot, the idea is to provide a (possibly unreachable) upper bound for each query [21, 22].

Letting $\mathrm{offset}(i)$ denote the offset position of $i \in M$, S is defined as:

$$S\text{-measure} = \frac{\sum_{i \in M} w(i) \max(0, L - \mathrm{offset}(i))}{\sum_{i \in I} w(i) \max(0, L - \mathrm{offset}^*(v(i)))}, \quad (6)$$

The original paper that proposed S used L = 1,000 [21], while 1CLICK-1 used L = 500 [22]. The former means that the user has about two minutes to examine the X-string, while the latter means that he only has one minute. A paper following up those two studies suggested that setting L to 250 and 500 yields different results, and is useful for evaluating systems for the 1CLICK task in different perspectives [20]. Therefore, we use $L = 250$ and $L = 500$ to evaluate submitted runs in 1CLICK-2.

For English subtasks, vital strings were used instead of iUnits to compute the S-measure. PMO for English subtasks is defined by sorting all vital strings by their weight (first key) and length (second key) as defined in the original S-measure definition.

5.3.3 T-measure

T-measure is a precision-like metric, which was introduced by Sakai and Kato [20] for distinguishing two X-strings that include the same iUnits at the same positions with irrelevant text of different lengths. T is defined as the faction of the sum of vital string lengths to the length of an X-string. Thus,

$$T\text{-measure} = \frac{\sum_{i \in M} |v(i)|}{|X|}, \quad (7)$$

where $|X|$ is the length of an X-string.

5.3.4 S♯-measure

S♯-measure is an F-measure-like combination of S and T, and was defined as follows:

$$S\sharp\text{-measure} = \frac{(1+\beta^2)TS}{\beta^2 T + S}. \quad (8)$$

Sakai and Kato [20] showed that $S\sharp$ with a heavy emphasis ($\beta = 10$) on S-measure can evaluates the terseness of an X-string, and yet achieve discriminative power that is comparable to S-measure. Therefore, we used S♯-measure as the primary metric in 1CLICK-2.

6. MAIN TASK RESULTS

We present a list of submitted runs and the official results in this section.

6.1 Submitted Runs

Tables 7 and 8 show runs submitted to English and Japanese Main tasks, respectively. The second column shows the SYSDESC (system description) field for each run (see Section 2.1). For Japanese Main tasks, we asked four assessors to generate an X-string for each query, and labeled the result as MANUAL. We expected that those MANUAL runs were approximations of realistic upper bound performance in the 1CLICK task. Note that MANUAL runs do not include any result for 27 queries listed in Appendix C, as we had generated those runs before the Japanese test queries were fixed.

6.2 Official Results of English Subtasks

This section reports the official English subtask results and analysis on the evaluation results.

We created four systems to provide baseline results using some simple heuristics. Three use no manual work but do rely on Microsoft Bing search result information, and the last uses ORACLE.

- **BASELINE-SNIPPETS** - Concatenates snippets from the search engine ranking file in rank order until the character limit is reached.
- **BASELINE-WIKI-HEAD** - Reads from the start of a cleaned version of the first Wikipedia page in the search results until the character limit is reached; returns empty if no Wikipedia article exists.
- **BASELINE-WIKI-KWD** - Takes a cleaned version of the first Wikipedia page in the search results. If the query has keywords not contained in the Wikipedia article title (as returned by the search engine results), concatenate sentences from the Wikipedia page which contain the greatest number of occurrences of those query terms not found in the title until the character limit is reached. Otherwise proceed as in BASELINE-WIKI-HEAD.
- **BASELINE-ORCL** - This utilizes the same nugget system as was used to extract candidate source sentences, however the only manual feedback used was the ORACLE documents provided to participants. No other feedback on documents or sentences was used. This is therefore somewhat closer to a participant run, but as no other ORACLE runs were submitted and given that there could be some bias introduced due to using the same system as the extraction system, it is included here as a BASELINE.

6.2.1 Results

In presenting the results, we break down the performance of the various systems using many factors in order to elucidate the benefits of certain participant methods and the difficulties of certain queries and query categories. First, in Figure 4 we see the $S\sharp$ score for all participant systems averaged across all queries.

We see here that overall many of the baseline systems performed the best. This is somewhat expected, and many of the queries were general enough to have full Wikipedia documents devoted to them, in which case fairly comprehensive summaries already existed on the internet. Specifically, note that the mobile summaries performed better then the desktop ones, indicating, again as expected, that the task becomes harder with more characters.

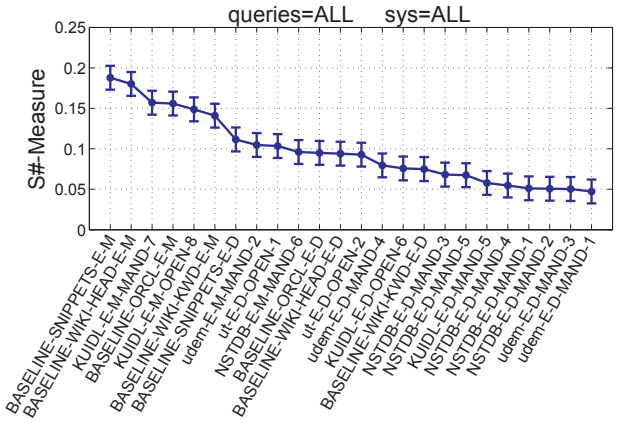


Figure 4: English Subtask: $S\sharp$ score for all systems averaged across all queries, with Tukey's HSD significance intervals.

As a result, it is easier to examine the overall rankings separately for Mobile and Desktop runs, as can be seen in Figure 5. We see that the BASELINE results using snippets dominates for both Desktop and Mobile runs, but the lead text from Wikipedia articles only leads for Mobile runs. A number of participant teams did well in relation to these runs, and as we will see, beat the baselines for certain types of queries. In order to examine this, we break down the averages by query category, summary type, and query type.

Additionally, we computed a variety of significance tests, including one-way ANOVA and Tukey's Honest Significant Difference tests. Given the number of summaries with no matches, there was little indicated significance based on variance. We discuss some reasons for this variance in Section 6.2.2. We see significance in terms of the difference in means and variance for a set sample size (as all runs were over the same number of queries) in Figure 4.

In Figure 6, we see the $S\sharp$ scores for each group averaged over groups of query categories. CELEBRITY queries consist of ACTOR, ATHLETE, ARTIST, and POLITICIAN queries; LOCATION queries consist of FACILITY and GEO queries; PRECISE queries consist of DEFINITION and QA queries.

For CELEBRITY queries, the baseline systems tend to outperform other systems for Mobile summaries, but for Desktop summaries, there are a number of systems which perform better. And for LOCATION queries, KUIDL is the best performing Mobile system and NSTDB the best non-Oracle Desktop system.

Another important breakdown concerns the general difficulty of the query. During query formulation, a worry was that the simple Wikipedia baseline would perform too well on CELEBRITY queries consisting of just a name due to the fact that well-written Wikipedia pages likely exist for all such queries. For that reason, specifiers were added to many of the queries, e.g. changing "Michael Jackson" to "Michael Jackson death." To an extent, this makes them more like QA queries, queries for which specific information is requested. As such, Figure 8 shows average scores for SPECIFIC vs NONSPECIFIC queries. The breakdown of the classifications of such queries can be seen in Table 21 in Appendix A.

These results show that there is certainly room for improvement in automatic summarization. It is also likely true that our assessors preferred more readable summaries (such as the baselines) and thus any assessor matching mistakes are more probable on crabbed summaries. Different summarizer systems should be used for different types of queries, especially to identify clean and concise ways in

Table 7: 1CLICK-2 runs submitted to English Main tasks. The run name is of a format "<team>-<lang>-<device>-<source>-<priority>".

run name	SYSDESC field (line 1 of the run file)
KUIDL-E-D-MAND-5	IE based on HTML structure & Web search snippet summarization
KUIDL-E-D-OPEN-6	Web-search-based query classification & IE based on HTML structure & Web search snippet summarization
KUIDL-E-M-MAND-7	IE based on HTML structure & Web search snippet summarization
KUIDL-E-M-OPEN-8	Web-search-based query classification & IE based on HTML structure & Web search snippet summarization
NSTDB-E-D-MAND-1	EF: extracting non-overlapped elements in a document by element-score-order (INEX's Focused task in Ad hoc track)
NSTDB-E-D-MAND-2	ER: extracting non-overlapped elements with grouped by document (INEX's Relevant in context task in Ad hoc track)
NSTDB-E-D-MAND-3	DR: extracting non-overlapped elements with grouped by document and ordered by documents' score (INEX's Relevant in context task in Ad hoc track)
NSTDB-E-D-MAND-4	EB: extracting only one element per document by element-score-order (INEX's Best in Context task in Ad hoc track)
NSTDB-E-D-MAND-5	DB: extracting only one element per document by document-score-order (INEX's Best in Context task in Ad hoc track)
NSTDB-E-M-MAND-6	mobileEF: extracting non-overlapped elements in a document by element-score-order for mobile (INEX's Focused task in Ad hoc track)
NUIR-E-D-MAND-1	Concatenates snippets from top documents.
NUIR-E-M-MAND-2	Concatenates snippets from top documents.
NUIR-E-D-MAND-3	Takes first string of text from first Wikipedia document.
NUIR-E-M-MAND-4	Takes first string of text from first Wikipedia document.
NUIR-E-D-MAND-5	Takes wikipedia text with query terms.
NUIR-E-M-MAND-6	Takes wikipedia text with query terms.
NUIR-E-D-MAND-7	Pseudo-relevance document feedback and mutual nugget-document reinforcement.
NUIR-E-M-MAND-8	Pseudo-relevance document feedback and mutual nugget-document reinforcement.
udem-E-D-MAND-1	ILP Hunter-Gatherer – desktop
udem-E-M-MAND-2	ILP Hunter-Gatherer – mobile
udem-E-D-MAND-3	"wiki based pattern extraction + learning nugget weight + ILP" – desktop
udem-E-D-MAND-4	Basic Hunter-Gatherer – desktop
ut-E-D-OPEN-1	API-based Information Extraction System without partial matching
ut-E-D-OPEN-2	API-based Information Extraction System with partial matching

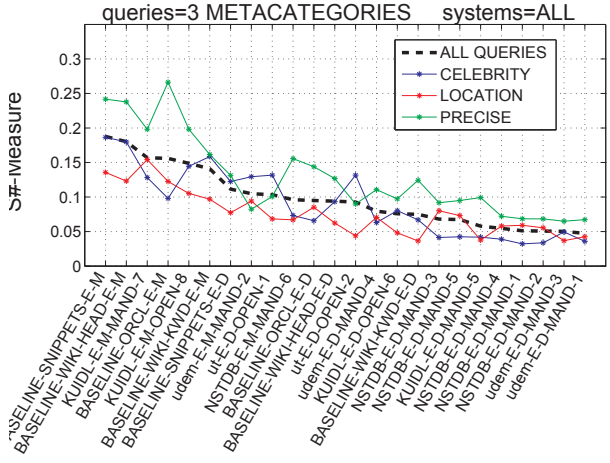


Figure 6: English Subtask: $S\sharp$ score for all systems averaged across each query Meta-category. CELEBRITY={ACTOR,ATHLETE,ARTIST,POLITICIAN}; LOCATION={GEO,FACILITY}; PRECISE ={DEFINITION,QA}.

which to deliver the content to users to minimize reading time.

6.2.2 *Analysis*

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.

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 Âů HP Pavilion Âů Honda Center
Âů Jobing.com Arena Âů Staples Center"],
["Broke ground":  "April 29, 1993"],
["Culture":  "Celtics parquet floor Celtic
```

Table 8: 1CLICK-2 runs submitted to Japanese Main tasks. The filename is of a format "<team>-<lang>-<device>-<source>-<priority>".

run name	SYSDESC field (line 1 of the run file)
KUIDL-J-D-MAND-1	IE based on HTML structure & Web search snippet summarization
KUIDL-J-D-OPEN-2	Web-search-based query classification & IE based on HTML structure & Web search snippet summarization
KUIDL-J-M-MAND-3	IE based on HTML structure & Web search snippet summarization
KUIDL-J-M-OPEN-4	Web-search-based query classification & IE based on HTML structure & Web search snippet summarization
MANUAL-J-D-OPEN-1	Manual runs
MANUAL-J-D-OPEN-2	Manual runs
MANUAL-J-D-OPEN-3	Manual runs
MANUAL-J-D-OPEN-4	Manual runs
MSRA-J-D-MAND-1	query type classification; attributes with tf-idf weights
MSRA-J-D-MAND-2	query type classification; attributes with binary weights
MSRA-J-D-MAND-3	query type classification; no attributes
MSRA-J-D-MAND-4	no query type classification
ORG-J-D-MAND-1	top-ranked snippets
ORG-J-D-MAND-2	first sentences in Wikipedia
ORG-J-D-ORCL-3	top-ranked snippets
TTOKU-J-M-ORCL-1	A query-oriented summarization system based on subtree extraction based on the query relevance score.
TTOKU-J-D-ORCL-2	A query-oriented summarization system based on subtree extraction based on the query relevance score.
TTOKU-J-M-MAND-3	A query-oriented summarization system based on subtree extraction based on the query relevance score.

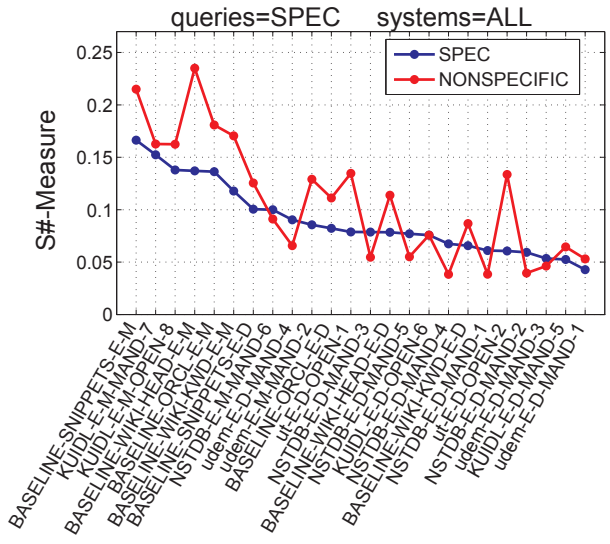


Figure 7: English Subtask: $S\sharp$ score for all systems averaged across PRECISE queries.

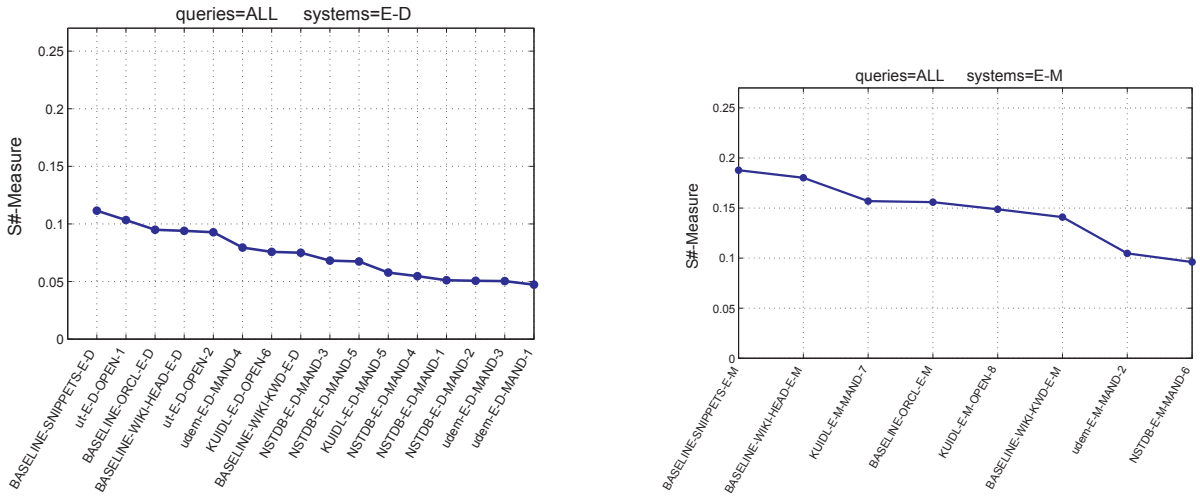

```
Pride Greatest game ever played Tommy
Points "Love ya, Cooz!" Close, but no
cigar!  Bill Russell Beat L.A. Mike Gorman
Johnny Most "Havlicek stole the ball!"
Henderson steals the ball!  Bird steals
the ball!  Boston Garden North Station
The Heat Game Memorial Day Massacre Larry
Legend DJ The Chief The Truth The Big Three
The Big Four"], ["Opened":  "September
30, 1995"], ["Architect":  "Ellerbe
Becket, Inc.[2]"], ["Coordinates":  "42°
21!58.69"N 71° 3!44.02"WâĂŃ / âĂŃ42.3663028
° N 71.0622278° WâĂŃ / 42.3663028;
-71.0622278Coordinates:  42° 21!58.69"N 71
° 3!44.02"WâĂŃ / âĂŃ42.3663028° N 71.0622278
° WâĂŃ / 42.3663028; -71.0622278"], ["NBA
Championships (17)":  "1957 1959 1960
19611962196319641965196619681968familyn 1969
1974 1976 1981 1984 1986 2008"], ["Retired
numbers":  "00 1 2 3 6 10 14 15 16 17 18
LOSCY 19 21 22 23 24 25 31 32 33 35 MIC"],
["Operator":  "Delaware North Companies"]
```


Readability also explains why the baseline summaries scores so high: the baselines implement simple ideas on Wikipedia pages, and most of them consist of large chunks of Wikipedia text, making them quite readable.

Assessor quality and issues.

We recognize that the assessing effort for the 1CLICK English subtasks was a massive challenge, and, while a significant effort has been made both on the extraction and on matching, these assessments are incomplete at best; thus the results might be inconclusive. Here we point to some of the assessing problems and possible solutions for future efforts:

- Missing relevant information (nugget). Assessors sometimes lost patience with reading through non-relevant information and as a result they might have missed relevant pieces when they appeared. A better assessor is the one naturally trained



(a) Desktop summaries. (b) Mobile summaries.

Figure 5: English Subtask: $S\sharp$ score averaged across ALL queries, broken down by summary type.

to read text while looking for certain facts, such as people previously employed as analysts.

- Missing proper matches. Besides fatigue and interest in the topic, a good reason for missing a match is that the match is not complete, so the assessor has to make an internal determination on whether the text in the summary is "close enough" to match the iUnit. Other reasons include words that match in a secondary meaning (unfamiliar to the assessor), or unfamiliarity with certain English constructs (some of our assessor were not native English speakers). As a solution, better/previously trained assessors would do a better job.

- The task was harder than traditional document relevance judgment. Not surprisingly, the assessor effort (extraction and matching) was significantly harder than expressing graded opinion on documents, and in some cases random decisions were made in lack of better options. A possible remedy is to have multiple assessors perform the extraction and the matching, and to look for consensus through debate when in disagreement.

6.3 Official Results of Japanese Subtasks

Table 10 shows the official mean $S\sharp$, S, T-measure, and weighted recall performances over 100 Japanese queries for the submitted runs except MANUAL runs. The column **I** indicates that the score was computed based on the *intersection* between sets of iUnit matches by two assessors, while **U** indicates that the score was computed based on the *union* between sets of iUnit matches by two assessors. The offset of iUnit matches is defined as the minimum offset of iUnit matches by two assessors in both of the cases. The runs are ranked by the mean $S\sharp$-measure with **I**. Figure 9 visualizes the official mean $S\sharp$-measure performances ($L = 500$) shown in Table 10.

Table 11 shows the official mean $S\sharp$-measure performances per query type, and Figures 10 and 11 visualize the performances shown in the table.

It can be observed that the three "ORG" runs are the overall top performers in Table 10. These are actually simple baseline runs submitted by the organizers' 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: Table 12 shows p-values two-sided randomized Tukey's HSD in terms of S♯-measure performances over 100 Japanese queries ($L = 500$).

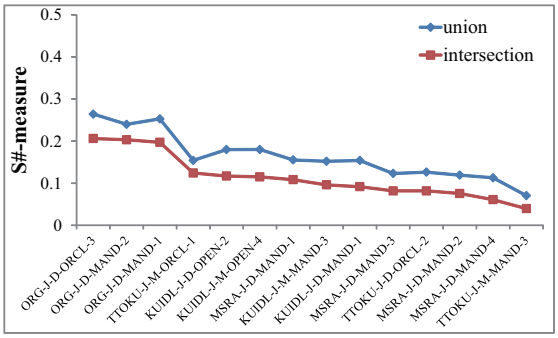


Figure 9: Japanese Subtask: Mean S♯-measure performances over 100 queries ($L = 500$). The x axis represents runs sorted by Mean S♯ with the intersection iUnit match data.

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, while they are not as effective for FACILITY, GEO and QA.

Table 13 shows mean $S\sharp$, S, T-measure, and weighted recall performances over the 73 queries for all runs including the MANUAL ones. Figure 12 shows the mean S♯-measure ($L = 500$) shown in the table. Recall that MANUAL runs generated X-strings only for those 73 queries. Table 14 and its graphs drawn in Figures 13 and 14 show the per-query performances over the 73 queries. It can be observed that three of the four MANUAL runs far outperform the submitted automatic runs. These three runs are statistically signifi-

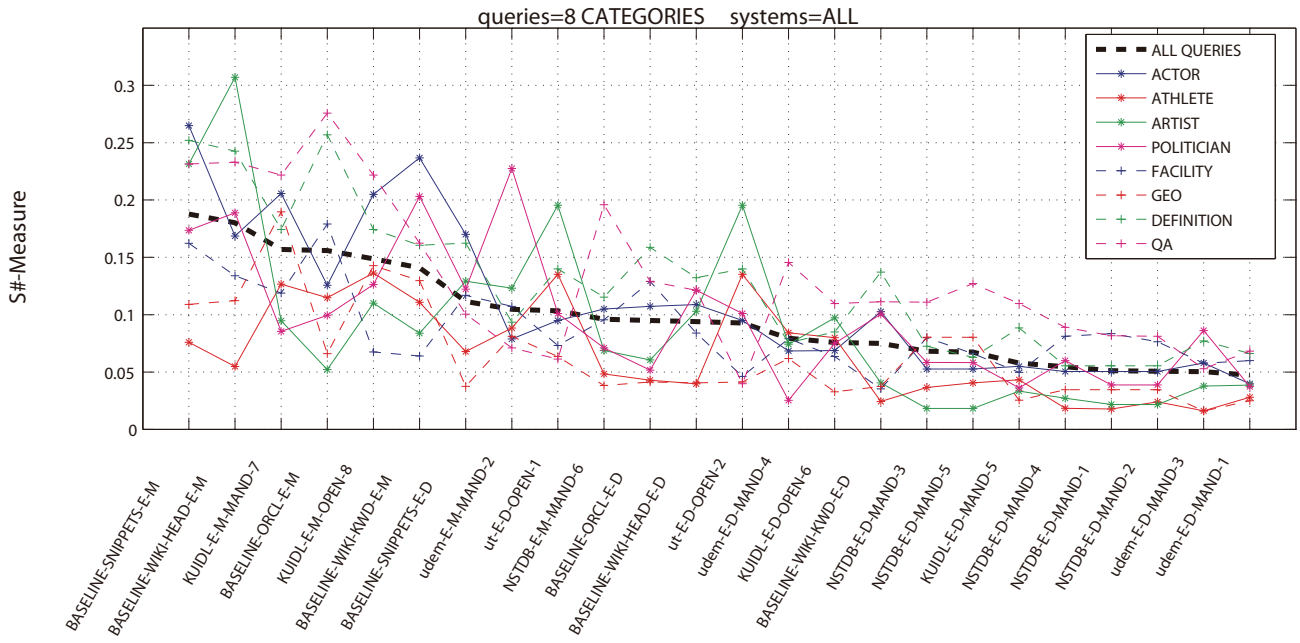


Figure 8: English Subtask: $S\sharp$ score for all systems averaged across each query category.

Table 9: English Subtask: $S\sharp$ score for all systems averaged across each query category. In parenthesis: the number of queries for which a summary was produced *and matched at least a vital string*.

		CELEBRITY				LOCATION		PRECISE	
SYS	ALL(100)	ACTOR(10)	ATHLE(10)	ARTIS(10)	POLIT(10)	FACIL(15)	GEO(15)	DEFIN(15)	QA(15)
BASELINE-SNIPPETS-E-M	0.188(66)	0.265(9)	0.076(5)	0.231(6)	0.174(7)	0.162(11)	0.109(7)	0.252(10)	0.231(11)
BASELINE-WIKI-HEAD-E-M	0.180(68)	0.169(9)	0.055(5)	0.307(9)	0.189(7)	0.134(10)	0.112(5)	0.243(12)	0.233(11)
KUIDL-E-M-MAND-7	0.157(61)	0.206(7)	0.127(6)	0.095(6)	0.085(6)	0.119(9)	0.190(8)	0.174(9)	0.222(10)
BASELINE-ORCL-E-M	0.156(62)	0.126(6)	0.115(5)	0.052(5)	0.100(3)	0.179(11)	0.066(5)	0.257(13)	0.276(14)
KUIDL-E-M-OPEN-8	0.149(55)	0.205(7)	0.136(6)	0.110(5)	0.126(6)	0.067(6)	0.143(6)	0.174(9)	0.222(10)
BASELINE-WIKI-KWD-E-M	0.141(55)	0.237(9)	0.111(4)	0.084(7)	0.203(5)	0.064(7)	0.130(4)	0.161(9)	0.162(10)
BASELINE-SNIPPETS-E-D	0.112(87)	0.170(10)	0.068(9)	0.129(10)	0.122(9)	0.117(15)	0.038(8)	0.162(14)	0.100(12)
udem-E-M-MAND-2	0.105(48)	0.079(5)	0.088(5)	0.123(8)	0.227(4)	0.107(7)	0.081(4)	0.094(8)	0.071(7)
ut-E-D-OPEN-1	0.103(73)	0.095(8)	0.135(5)	0.195(10)	0.101(6)	0.073(12)	0.064(9)	0.140(11)	0.061(12)
NSTDB-E-M-MAND-6	0.096(51)	0.105(6)	0.048(6)	0.069(3)	0.071(4)	0.096(10)	0.038(3)	0.115(8)	0.196(11)
BASELINE-ORCL-E-D	0.095(72)	0.107(9)	0.043(7)	0.060(6)	0.052(4)	0.129(12)	0.041(6)	0.159(14)	0.129(14)
BASELINE-WIKI-HEAD-E-D	0.094(74)	0.109(10)	0.040(6)	0.103(8)	0.121(8)	0.084(11)	0.041(6)	0.132(13)	0.121(12)
ut-E-D-OPEN-2	0.093(68)	0.095(8)	0.135(5)	0.195(10)	0.101(7)	0.046(11)	0.041(8)	0.140(11)	0.040(8)
udem-E-D-MAND-4	0.080(66)	0.068(7)	0.084(8)	0.074(6)	0.025(4)	0.079(12)	0.062(6)	0.076(10)	0.146(13)
KUIDL-E-D-OPEN-6	0.076(72)	0.069(7)	0.080(7)	0.097(6)	0.075(10)	0.064(11)	0.033(6)	0.085(11)	0.110(14)
BASELINE-WIKI-KWD-E-D	0.075(62)	0.103(10)	0.024(5)	0.041(7)	0.100(6)	0.035(7)	0.037(3)	0.137(13)	0.111(11)
NSTDB-E-D-MAND-3	0.068(55)	0.053(6)	0.037(6)	0.018(4)	0.058(5)	0.080(10)	0.080(6)	0.072(8)	0.111(10)
NSTDB-E-D-MAND-5	0.067(56)	0.053(6)	0.041(6)	0.018(4)	0.058(5)	0.066(10)	0.080(6)	0.063(8)	0.127(11)
KUIDL-E-D-MAND-5	0.058(63)	0.055(6)	0.043(5)	0.033(4)	0.036(4)	0.050(10)	0.025(8)	0.088(12)	0.110(14)
NSTDB-E-D-MAND-4	0.055(55)	0.050(7)	0.018(5)	0.027(4)	0.060(6)	0.081(10)	0.035(5)	0.055(7)	0.089(11)
NSTDB-E-D-MAND-1	0.051(54)	0.050(7)	0.018(5)	0.022(4)	0.039(5)	0.083(10)	0.035(5)	0.055(7)	0.082(11)
NSTDB-E-D-MAND-2	0.051(53)	0.050(7)	0.024(5)	0.022(4)	0.039(5)	0.076(10)	0.035(5)	0.055(7)	0.081(10)
udem-E-D-MAND-3	0.050(57)	0.058(7)	0.016(4)	0.038(6)	0.086(4)	0.058(11)	0.016(4)	0.077(11)	0.053(10)
udem-E-D-MAND-1	0.047(62)	0.040(7)	0.028(6)	0.039(7)	0.037(5)	0.060(12)	0.025(4)	0.066(11)	0.068(10)
AVERAGE $S\sharp$	0.098	0.109	0.066	0.091	0.095	0.088	0.065	0.126	0.131

Table 10: Japanese Subtask: Mean $S\sharp$, S, T-measure, and weighted recall over 100 Japanese queries. Runs sorted by the mean $S\sharp$-measure with I.

run name	$S\sharp$ ($L=500$)		$S\sharp$ ($L=250$)		S ($L=500$)		S ($L=250$)		T		Weighted recall	
	I	**U**	**I**	**U**	**I**	**U**	**I**	**U**	**I**	**U**	**I**	**U**
ORG-J-D-ORCL-3	0.206	0.264	0.227	0.275	0.210	0.268	0.233	0.280	0.096	0.141	0.157	0.216
ORG-J-D-MAND-2	0.203	0.240	0.233	0.267	0.206	0.242	0.237	0.271	0.127	0.163	0.146	0.173
ORG-J-D-MAND-1	0.197	0.253	0.209	0.248	0.201	0.256	0.214	0.252	0.099	0.144	0.164	0.226
TTOKU-J-M-ORCL-1	0.124	0.154	0.121	0.149	0.127	0.157	0.124	0.151	0.107	0.150	0.049	0.063
KUIDL-J-D-OPEN-2	0.117	0.180	0.133	0.188	0.119	0.182	0.136	0.192	0.057	0.105	0.088	0.146
KUIDL-J-M-OPEN-4	0.115	0.180	0.116	0.180	0.116	0.183	0.117	0.181	0.104	0.175	0.053	0.081
MSRA-J-D-MAND-1	0.108	0.155	0.123	0.168	0.110	0.157	0.125	0.171	0.067	0.108	0.082	0.126
KUIDL-J-M-MAND-3	0.096	0.152	0.098	0.153	0.099	0.156	0.098	0.154	0.085	0.145	0.047	0.073
KUIDL-J-D-MAND-1	0.092	0.154	0.095	0.151	0.093	0.156	0.098	0.153	0.052	0.098	0.068	0.127
MSRA-J-D-MAND-3	0.082	0.123	0.092	0.127	0.083	0.125	0.094	0.129	0.041	0.075	0.065	0.101
TTOKU-J-D-ORCL-2	0.082	0.126	0.091	0.132	0.083	0.128	0.094	0.135	0.050	0.087	0.059	0.108
MSRA-J-D-MAND-2	0.075	0.119	0.087	0.127	0.077	0.121	0.089	0.130	0.046	0.081	0.051	0.088
MSRA-J-D-MAND-4	0.061	0.113	0.065	0.111	0.062	0.114	0.067	0.115	0.036	0.078	0.051	0.104
TTOKU-J-M-MAND-3	0.040	0.070	0.038	0.066	0.041	0.072	0.040	0.068	0.037	0.071	0.018	0.031

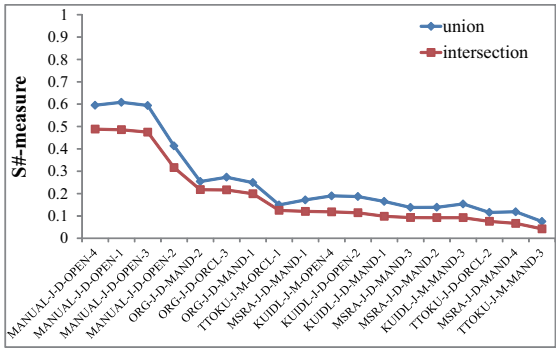


Figure 12: Japanese Subtask: Mean S♯-measure performances over 73 queries ($L=500$). The x axis represents runs sorted by Mean S♯ with the intersection iUnit match data.

cantly better than the other runs, and are statistically indistinguishable from one another: Table 15 shows p-values two-sided randomized Tukey's HSD in terms of S♯-measure performances over 73 Japanese queries ($L=500$). 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.

Table 16 shows the mean of the sum of two assessors' readability and trustworthiness scores. Recall that the organizers asked assessors to evaluate those two criteria on a four point scale: very low (-2), low (-1), high ($+1$), and very high ($+2$).

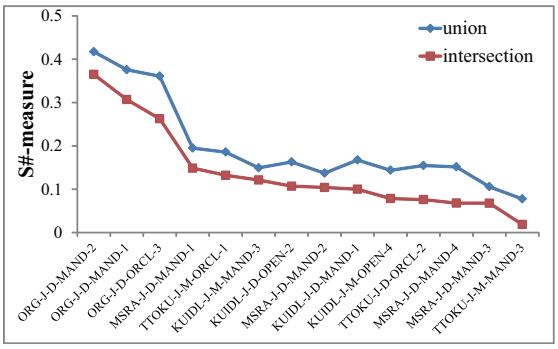
Table 11: Japanese Subtask: Mean per-query $S\sharp$-measure over 100 Japanese queries ($L = 500$). Bold font indicates the highest performance in each row.

run name	ARTIST		ACTOR		POLITICIAN		ATHLETE	
	I	**U**	**I**	**U**	**I**	**U**	**I**	**U**
ORG-J-D-ORCL-3	0.263	0.361	**0.293**	**0.389**	**0.317**	**0.383**	0.306	0.336
ORG-J-D-MAND-2	**0.365**	**0.417**	0.291	0.370	0.273	0.320	**0.378**	**0.449**
ORG-J-D-MAND-1	0.307	0.376	0.279	0.378	0.241	0.330	0.298	0.352
TTOKU-J-M-ORCL-1	0.132	0.186	0.160	0.181	0.096	0.144	0.109	0.150
KUIDL-J-D-OPEN-2	0.107	0.163	0.130	0.217	0.124	0.178	0.230	0.278
KUIDL-J-M-OPEN-4	0.078	0.144	0.140	0.247	0.067	0.130	0.199	0.254
MSRA-J-D-MAND-1	0.149	0.195	0.165	0.226	0.138	0.161	0.150	0.254
KUIDL-J-M-MAND-3	0.121	0.149	0.150	0.237	0.036	0.150	0.171	0.187
KUIDL-J-D-MAND-1	0.100	0.167	0.115	0.175	0.056	0.111	0.173	0.213
MSRA-J-D-MAND-3	0.068	0.106	0.096	0.148	0.037	0.059	0.107	0.190
TTOKU-J-D-ORCL-2	0.076	0.155	0.133	0.202	0.098	0.149	0.081	0.111
MSRA-J-D-MAND-2	0.104	0.137	0.114	0.184	0.085	0.134	0.138	0.215
MSRA-J-D-MAND-4	0.068	0.151	0.069	0.122	0.029	0.065	0.042	0.125
TTOKU-J-M-MAND-3	0.019	0.078	0.071	0.106	0.043	0.076	0.017	0.076

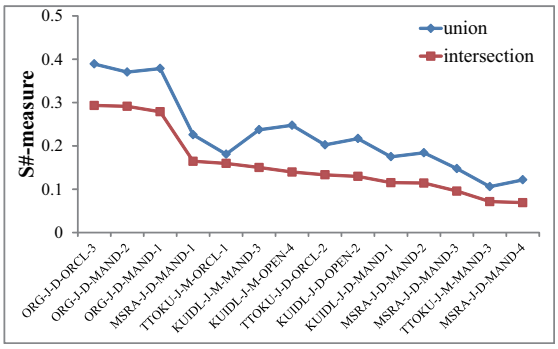
run name	FACILITY		GEO		DEFINITION		QA	
	I	**U**	**I**	**U**	**I**	**U**	**I**	**U**
ORG-J-D-ORCL-3	0.097	0.130	0.069	0.109	0.294	0.382	0.127	0.161
ORG-J-D-MAND-2	0.057	0.060	0.035	0.058	0.312	0.350	0.079	0.094
ORG-J-D-MAND-1	0.051	0.100	0.084	0.108	**0.350**	**0.397**	0.080	0.123
TTOKU-J-M-ORCL-1	0.060	0.105	**0.177**	**0.198**	0.036	0.050	**0.223**	**0.233**
KUIDL-J-D-OPEN-2	0.182	0.263	0.009	0.030	0.128	0.220	0.067	0.128
KUIDL-J-M-OPEN-4	**0.224**	**0.296**	0.013	0.060	0.154	0.213	0.051	0.114
MSRA-J-D-MAND-1	0.140	0.182	0.060	0.138	0.100	0.123	0.021	0.034
KUIDL-J-M-MAND-3	0.136	0.218	0.013	0.033	0.089	0.152	0.082	0.127
KUIDL-J-D-MAND-1	0.121	0.204	0.011	0.039	0.114	0.183	0.069	0.156
MSRA-J-D-MAND-3	0.143	0.175	0.054	0.090	0.116	0.183	0.026	0.037
TTOKU-J-D-ORCL-2	0.086	0.151	0.075	0.099	0.026	0.036	0.098	0.143
MSRA-J-D-MAND-2	0.089	0.151	0.029	0.063	0.072	0.111	0.018	0.023
MSRA-J-D-MAND-4	0.059	0.096	0.106	0.165	0.076	0.142	0.025	0.040
TTOKU-J-M-MAND-3	0.000	0.011	0.054	0.071	0.018	0.044	0.092	0.119

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

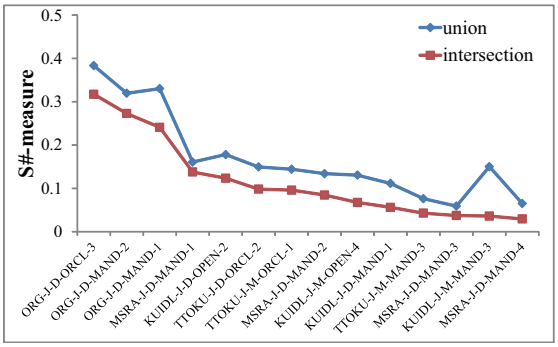
		KUIDL				MSRA				ORG			TTOKU		
		1	2	3	4	1	2	3	4	1	2	3	1	2	3
KUIDL	1		0.997	1.000	0.998	1.000	1.000	1.000	0.969	**0.000**	**0.000**	**0.000**	0.952	1.000	0.378
	2			1.000	1.000	1.000	0.749	0.909	0.254	**0.007**	**0.002**	**0.001**	1.000	0.909	**0.010**
	3				1.000	1.000	1.000	1.000	0.910	**0.000**	**0.000**	**0.000**	0.987	1.000	0.245
	4					1.000	0.810	0.943	0.311	**0.005**	**0.001**	**0.001**	1.000	0.942	**0.017**
MSRA	1						0.949	0.994	0.541	**0.001**	**0.000**	**0.000**	1.000	0.994	**0.047**
	2							1.000	1.000	**0.000**	**0.000**	**0.000**	0.491	1.000	0.896
	3								1.000	**0.000**	**0.000**	**0.000**	0.716	1.000	0.729
	4									**0.000**	**0.000**	**0.000**	0.103	1.000	0.999
ORG	1										1.000	1.000	**0.025**	**0.000**	**0.000**
	2											1.000	**0.009**	**0.000**	**0.000**
	3												**0.005**	**0.000**	**0.000**
TTOKU	1													0.711	**0.003**
	2														0.731
	3														



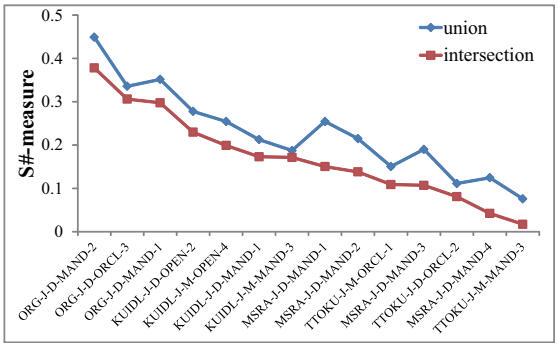
(a) ARTIST



(d) ACTOR

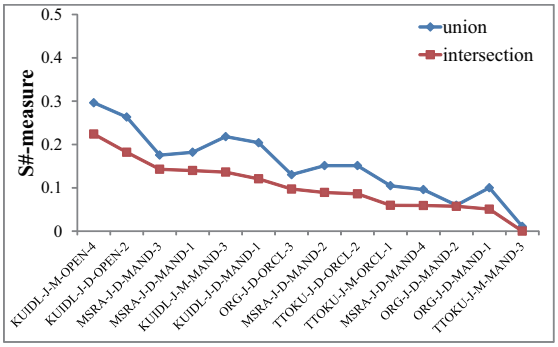


(b) POLITICIAN

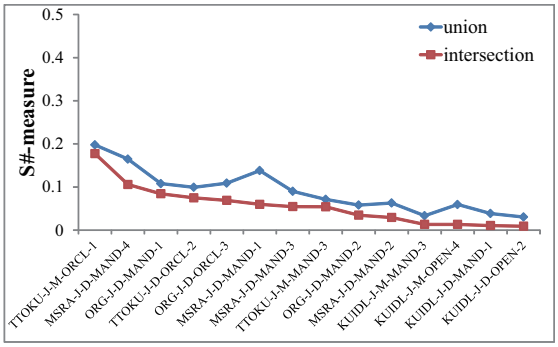


(e) ATHLETE

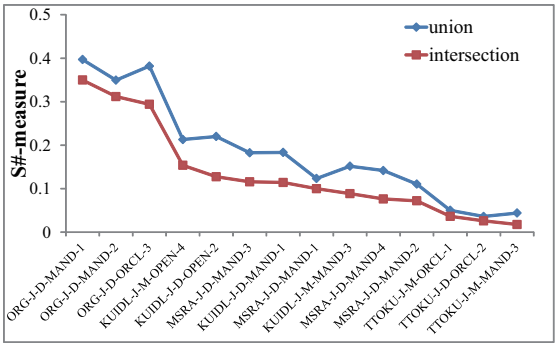
Figure 10: Japanese Subtask: Mean S♯-measure performances over 100 Japanese queries ($L = 500$) for ARTIST, ACTOR, POLITICIAN, and ATHLETE query types. The x axis represents runs sorted by Mean S♯ with the intersection iUnit match data.



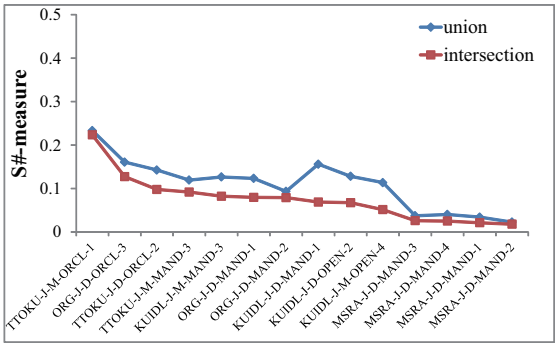
(c) FACILITY



(f) GEO



(f) DEFINITION

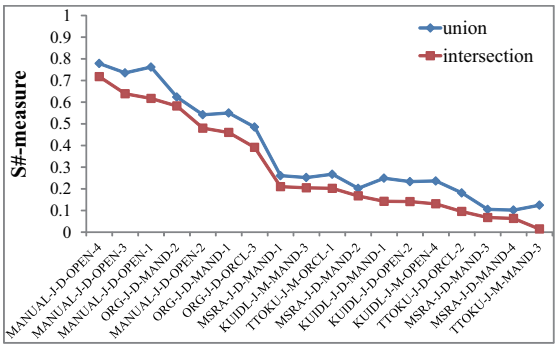


(f) QA

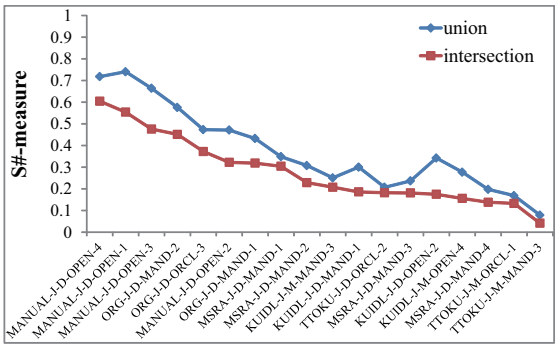
Figure 11: Japanese Subtask: Mean S♯-measure performances over 100 Japanese queries ($L = 500$) for FACILITY, GEO, DEFINITION, and QA query types. The x axis represents runs sorted by Mean S♯ with the intersection iUnit match data.

Table 13: Japanese Subtask: Mean $S\sharp$, S, T-measure, and weighted recall over 73 Japanese queries. Runs sorted by the mean $S\sharp$-measure with I.

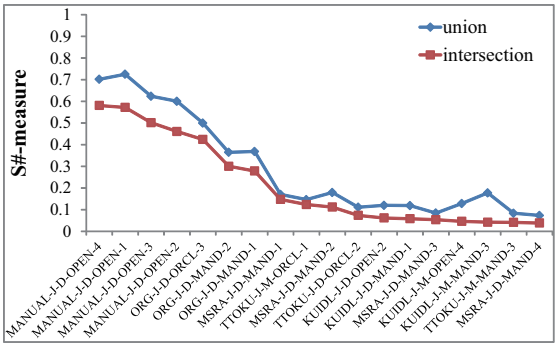
run name	$S\sharp$ ($L=500$)		$S\sharp$ ($L=250$)		S ($L=500$)		S ($L=250$)		T		Weighted recall	
	I	**U**	**I**	**U**	**I**	**U**	**I**	**U**	**I**	**U**	**I**	**U**
MANUAL-J-D-OPEN-4	0.488	0.595	0.534	0.616	0.494	0.604	0.560	0.651	0.380	0.521	0.359	0.465
MANUAL-J-D-OPEN-1	0.486	0.608	0.555	0.656	0.488	0.610	0.566	0.674	0.403	0.562	0.342	0.447
MANUAL-J-D-OPEN-3	0.475	0.594	0.524	0.627	0.478	0.598	0.538	0.650	0.362	0.498	0.342	0.441
MANUAL-J-D-OPEN-2	0.316	0.413	0.358	0.449	0.318	0.415	0.362	0.453	0.241	0.353	0.233	0.316
ORG-J-D-MAND-2	0.218	0.254	0.257	0.293	0.220	0.256	0.261	0.296	0.139	0.171	0.149	0.175
ORG-J-D-ORCL-3	0.216	0.273	0.250	0.300	0.220	0.277	0.257	0.306	0.101	0.149	0.151	0.204
ORG-J-D-MAND-1	0.199	0.249	0.217	0.256	0.202	0.252	0.222	0.260	0.104	0.147	0.155	0.202
TTOKU-J-M-ORCL-1	0.125	0.149	0.124	0.146	0.128	0.153	0.128	0.150	0.109	0.139	0.042	0.051
MSRA-J-D-MAND-1	0.121	0.172	0.141	0.191	0.122	0.174	0.144	0.195	0.071	0.114	0.083	0.128
KUIDL-J-M-OPEN-4	0.118	0.190	0.120	0.190	0.119	0.191	0.121	0.191	0.102	0.174	0.051	0.079
KUIDL-J-D-OPEN-2	0.114	0.187	0.132	0.196	0.116	0.189	0.136	0.200	0.056	0.114	0.081	0.146
KUIDL-J-D-MAND-1	0.099	0.165	0.103	0.165	0.100	0.167	0.105	0.167	0.060	0.108	0.071	0.128
MSRA-J-D-MAND-3	0.093	0.138	0.108	0.148	0.094	0.140	0.111	0.151	0.047	0.080	0.068	0.108
MSRA-J-D-MAND-2	0.092	0.139	0.108	0.153	0.094	0.141	0.111	0.157	0.054	0.089	0.062	0.102
KUIDL-J-M-MAND-3	0.092	0.154	0.094	0.156	0.093	0.154	0.095	0.157	0.074	0.138	0.035	0.063
TTOKU-J-D-ORCL-2	0.076	0.116	0.088	0.126	0.077	0.117	0.091	0.128	0.052	0.088	0.052	0.096
MSRA-J-D-MAND-4	0.067	0.119	0.076	0.126	0.068	0.120	0.078	0.130	0.039	0.079	0.049	0.093
TTOKU-J-M-MAND-3	0.042	0.075	0.041	0.073	0.044	0.077	0.044	0.076	0.033	0.069	0.014	0.027



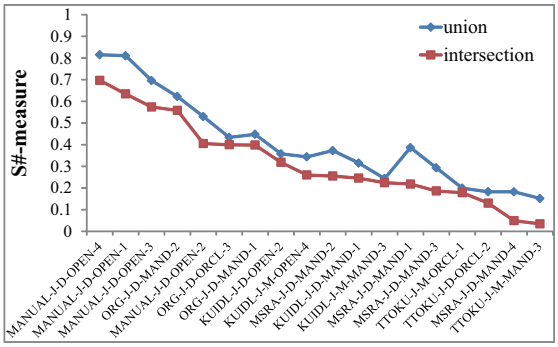
(a) ARTIST



(d) ACTOR



(b) POLITICIAN



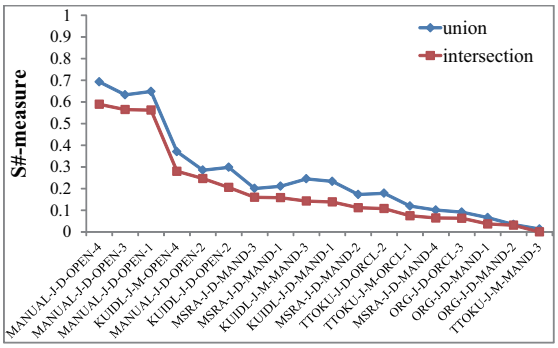
(e) ATHLETE

Figure 13: Japanese Subtask: Mean S♯-measure performances over 73 Japanese queries ($L=500$) for ARTIST, ACTOR, POLITICIAN, and ATHLETE query types. The x axis represents runs sorted by Mean S♯ with the intersection iUnit match data.

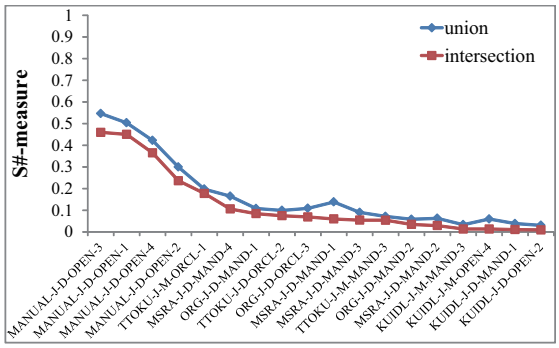
Table 14: Japanese Subtask: Mean per-query $S\sharp$-measure over 73 Japanese queries ($L = 500$). Bold font indicates the highest performance in each row.

run name	ARTIST		ACTOR		POLITICIAN		ATHLETE	
	I	**U**	**I**	**U**	**I**	**U**	**I**	**U**
MANUAL-J-D-OPEN-4	**0.718**	**0.779**	**0.605**	0.718	**0.581**	0.702	**0.697**	**0.815**
MANUAL-J-D-OPEN-1	0.617	0.762	0.554	**0.741**	0.572	**0.725**	0.634	0.811
MANUAL-J-D-OPEN-3	0.639	0.735	0.476	0.664	0.502	0.624	0.574	0.696
MANUAL-J-D-OPEN-2	0.480	0.542	0.322	0.471	0.461	0.600	0.405	0.530
ORG-J-D-MAND-2	0.582	0.624	0.451	0.576	0.300	0.365	0.558	0.622
ORG-J-D-ORCL-3	0.391	0.485	0.372	0.473	0.424	0.500	0.399	0.434
ORG-J-D-MAND-1	0.460	0.550	0.319	0.432	0.279	0.369	0.398	0.448
TTOKU-J-M-ORCL-1	0.202	0.267	0.133	0.168	0.123	0.146	0.178	0.199
MSRA-J-D-MAND-1	0.210	0.261	0.304	0.348	0.147	0.170	0.218	0.387
KUIDL-J-M-OPEN-4	0.131	0.236	0.156	0.277	0.046	0.128	0.260	0.344
KUIDL-J-D-OPEN-2	0.141	0.233	0.175	0.342	0.062	0.120	0.319	0.358
KUIDL-J-D-MAND-1	0.142	0.249	0.185	0.300	0.058	0.119	0.245	0.315
MSRA-J-D-MAND-3	0.067	0.105	0.181	0.237	0.053	0.084	0.186	0.293
MSRA-J-D-MAND-2	0.167	0.202	0.228	0.308	0.112	0.179	0.255	0.373
KUIDL-J-M-MAND-3	0.204	0.252	0.207	0.251	0.042	0.177	0.224	0.244
TTOKU-J-D-ORCL-2	0.095	0.181	0.182	0.207	0.074	0.111	0.130	0.183
MSRA-J-D-MAND-4	0.063	0.102	0.138	0.198	0.038	0.073	0.049	0.182
TTOKU-J-M-MAND-3	0.014	0.124	0.041	0.079	0.041	0.084	0.034	0.152

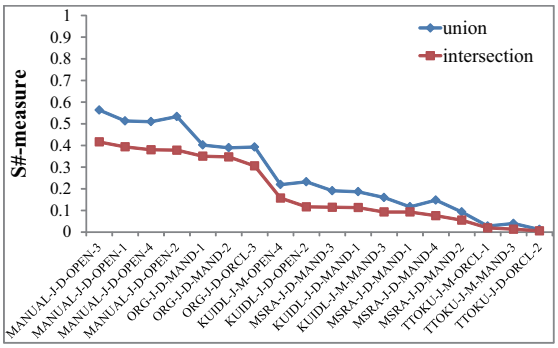
run name	FACILITY		GEO		DEFINITION		QA	
	I	**U**	**I**	**U**	**I**	**U**	**I**	**U**
MANUAL-J-D-OPEN-4	**0.589**	**0.693**	0.365	0.422	0.380	0.510	**0.364**	0.518
MANUAL-J-D-OPEN-1	0.562	0.648	0.450	0.504	0.393	0.513	0.350	**0.523**
MANUAL-J-D-OPEN-3	0.565	0.633	**0.460**	**0.547**	**0.416**	**0.563**	0.337	0.497
MANUAL-J-D-OPEN-2	0.246	0.285	0.236	0.299	0.378	0.533	0.231	0.330
ORG-J-D-MAND-2	0.031	0.034	0.035	0.058	0.347	0.389	0.064	0.079
ORG-J-D-ORCL-3	0.063	0.091	0.069	0.109	0.305	0.393	0.130	0.171
ORG-J-D-MAND-1	0.037	0.066	0.084	0.108	0.350	0.402	0.066	0.102
TTOKU-J-M-ORCL-1	0.074	0.119	0.177	0.198	0.019	0.028	0.159	0.165
MSRA-J-D-MAND-1	0.158	0.211	0.060	0.138	0.093	0.117	0.017	0.031
KUIDL-J-M-OPEN-4	0.280	0.370	0.013	0.060	0.157	0.219	0.011	0.059
KUIDL-J-D-OPEN-2	0.205	0.298	0.009	0.030	0.116	0.232	0.061	0.110
KUIDL-J-D-MAND-1	0.138	0.233	0.011	0.039	0.113	0.186	0.062	0.106
MSRA-J-D-MAND-3	0.160	0.201	0.054	0.090	0.115	0.191	0.009	0.023
MSRA-J-D-MAND-2	0.112	0.173	0.029	0.063	0.055	0.093	0.022	0.028
KUIDL-J-M-MAND-3	0.142	0.245	0.013	0.033	0.093	0.159	0.020	0.075
TTOKU-J-D-ORCL-2	0.108	0.178	0.075	0.099	0.006	0.012	0.042	0.087
MSRA-J-D-MAND-4	0.064	0.101	0.106	0.165	0.076	0.147	0.008	0.026
TTOKU-J-M-MAND-3	0.000	0.013	0.054	0.071	0.013	0.040	0.115	0.120



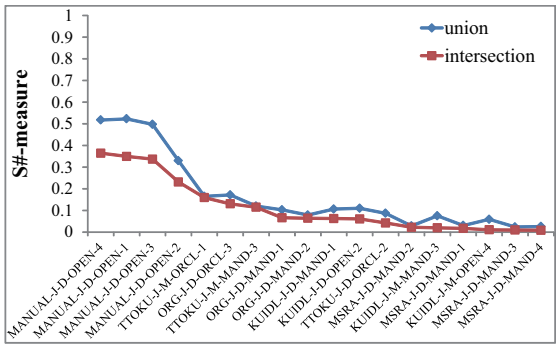
(c) FACILITY



(f) GEO



(f) DEFINITION



(f) QA

Figure 14: Japanese Subtask: Mean S♯-measure performances over 73 Japanese queries ($L = 500$) for FACILITY, GEO, DEFINITION, and QA query types. The x axis represents runs sorted by Mean S♯ with the intersection iUnit match data.

Table 15: Japanese Subtask: p-values of two-sided randomized Tukey's HSD in terms of S♯-measure performances over 73 Japanese queries ($L = 500$). Bold font indicates p-values $< \alpha = 0.05$.

		KUIDL				MANUAL				MSRA				ORG			TTOKU		
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	1	2	3
KUIDL	1		1.000	1.000	1.000	**0.000**	**0.000**	**0.000**	**0.000**	1.000	1.000	1.000	1.000	0.252	0.058	0.063	1.000	1.000	0.977
	2			1.000	1.000	**0.000**	**0.000**	**0.000**	**0.000**	1.000	1.000	1.000	0.997	0.562	0.208	0.223	1.000	1.000	0.827
	3				1.000	**0.000**	**0.000**	**0.000**	**0.000**	1.000	1.000	1.000	1.000	0.164	**0.032**	**0.036**	1.000	1.000	0.995
	4					**0.000**	**0.000**	**0.000**	**0.000**	1.000	1.000	1.000	0.994	0.660	0.273	0.294	1.000	0.999	0.763
MANUAL	1						**0.000**	1.000	1.000	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
	2							**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.069	0.289	0.268	**0.000**	**0.000**	**0.000**
	3								1.000	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
	4									**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
MSRA	1										1.000	1.000	0.988	0.711	0.321	0.339	1.000	0.998	0.719
	2											1.000	1.000	0.165	**0.032**	**0.036**	1.000	1.000	0.995
	3												1.000	0.169	**0.033**	**0.037**	1.000	1.000	0.995
	4													**0.012**	**0.001**	**0.001**	0.969	1.000	1.000
ORG	1														1.000	1.000	0.790	**0.039**	**0.000**
	2															1.000	0.406	**0.003**	**0.000**
	3																0.431	**0.003**	**0.000**
TTOKU	1																	0.995	0.618
	2																		1.000
	3																		

Table 16: Japanese Subtask: Mean of the sum of two assessors' readability and trustworthiness scores.

run name	readability	trustworthiness
KUIDL-J-D-MAND-1.tsv	-1.3	-1.48
KUIDL-J-D-OPEN-2.tsv	-1.32	-1.27
KUIDL-J-M-MAND-3.tsv	-1.23	-2.42
KUIDL-J-M-OPEN-4.tsv	-1.67	-2.46
MANUAL-J-D-OPEN-1.tsv	0.92	1.08
MANUAL-J-D-OPEN-2.tsv	0.54	0.74
MANUAL-J-D-OPEN-3.tsv	0.92	0.96
MANUAL-J-D-OPEN-4.tsv	0.57	0.91
MSRA-J-D-MAND-1.tsv	-1.43	-1.47
MSRA-J-D-MAND-2.tsv	-1.8	-1.97
MSRA-J-D-MAND-3.tsv	-1.89	-2.08
MSRA-J-D-MAND-4.tsv	-2.14	-2.16
ORG-J-D-MAND-1.tsv	-0.26	-0.13
ORG-J-D-MAND-2.tsv	1.5	0.82
ORG-J-D-ORCL-3.tsv	-0.31	0.09
TTOKU-J-D-ORCL-2.tsv	-2.35	-2.27
TTOKU-J-M-MAND-3.tsv	-2.68	-3.23
TTOKU-J-M-ORCL-1.tsv	-1.55	-2.28

7. QUERY CLASSIFICATION SUBTASK RESULTS

Table 17 shows submitted runs to the query classification subtask. Tables 18 and 20 show the official English and Japanese query classification subtask results, where the number of true positive (TP) and false positive (FP) are shown for each query type. It can be observed that it is difficult to identify DEFINITION type queries, while celebrity query types (i.e. ARTIST, ACTOR, POLITICIAN, and ATHLETE) are distinguished correctly.

Table 17: List of query classification subtask runs. The run name is of a format "<team>-QC-<priority>".

language	run name
English	KUIDL-QC-3
	KUIDL-QC-4
	NUTKS-QC-1
	NUTKS-QC-11
	NUTKS-QC-3
	NUTKS-QC-5
	NUTKS-QC-7
	NUTKS-QC-9
	ut-QC-1
	ut-QC-2
Japanese	HUKB-QC-1
	HUKB-QC-2
	KUIDL-QC-1
	KUIDL-QC-2
	MSRA-QC-1
	NUTKS-QC-10
	NUTKS-QC-12
	NUTKS-QC-2
	NUTKS-QC-4
	NUTKS-QC-6
	NUTKS-QC-8

Table 18: Official English query classification subtask results. The number of true positive (TP) and false positive (FP) are shown for each query type. Runs sorted by the accuracy, which is defined as the number of TP divided by the number of queries.

run name	ARTIST		ACTOR		POLITICIAN		ATHLETE		FACILITY		GEO		DEFINITION		QA		accuracy
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	
NUTKS-QC-3	10	0	9	0	10	3	10	0	12	1	13	2	13	2	12	3	0.89
NUTKS-QC-7	10	1	8	0	10	3	10	0	11	1	13	3	13	3	11	3	0.86
NUTKS-QC-1	9	0	9	1	10	2	10	1	12	2	11	3	10	2	13	5	0.84
NUTKS-QC-11	9	1	8	1	10	1	10	0	10	3	11	8	14	2	12	0	0.84
NUTKS-QC-9	9	1	8	1	9	2	9	0	9	5	10	8	14	3	12	0	0.80
NUTKS-QC-5	9	1	8	1	10	2	9	0	10	4	10	6	10	3	11	6	0.77
KUIDL-QC-3	9	3	4	2	6	0	7	0	6	4	9	7	13	15	14	1	0.68
ut-QC-1	9	7	5	2	6	2	7	1	6	4	6	9	12	20	4	0	0.55
ut-QC-2	9	7	5	1	5	3	7	3	5	3	8	17	12	12	3	0	0.54
KUIDL-QC-4	5	7	2	1	0	0	0	1	2	8	2	5	14	38	14	1	0.39

Table 19: Additional English Query Classification results including accuracy of nonspecific queries, and true positive and false positives for the meta-categories.

run name	NONSPEC accuracy	CELEBRITY		LOCATION		PRECISE	
		TP	FP	TP	FP	TP	FP
NUTKS-QC-3	0.9318	40	2	28	0	28	2
NUTKS-QC-7	0.9091	40	2	28	0	28	2
NUTKS-QC-11	0.8864	40	0	28	4	26	2
NUTKS-QC-1	0.8636	39	3	27	1	27	3
NUTKS-QC-9	0.8182	39	0	27	5	26	3
NUTKS-QC-5	0.8182	38	2	28	2	26	4
KUIDL-QC-3	0.7955	30	1	23	3	28	15
ut-QC-1	0.7045	32	7	17	8	26	10
ut-QC-2	0.6818	31	9	20	13	21	6
KUIDL-QC-4	0.4773	15	1	9	8	29	38

Table 20: Official Japanese query classification subtask results. The number of true positive (TP) and false positive (FP) are shown for each query type. Runs sorted by the accuracy, which is defined as the number of TP divided by the number of queries.

run name	ARTIST		ACTOR		POLITICIAN		ATHLETE		FACILITY		GEO		DEFINITION		QA		accuracy
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	
KUIDL-QC-1	9	2	10	2	10	0	9	0	12	4	15	2	10	3	12	0	0.87
NUTKS-QC-2	10	1	10	0	9	2	10	0	11	4	12	2	11	4	13	1	0.86
NUTKS-QC-10	10	1	10	1	9	1	9	0	13	5	12	1	10	3	13	2	0.86
NUTKS-QC-12	9	2	9	2	10	0	9	0	11	5	14	4	10	0	13	2	0.85
NUTKS-QC-4	9	3	10	1	10	0	9	0	11	3	13	4	11	4	11	1	0.84
NUTKS-QC-6	9	1	10	0	9	2	10	0	11	4	12	2	11	6	12	1	0.84
MSRA-QC-1	9	4	9	1	8	0	10	1	7	1	14	2	12	7	14	1	0.83
NUTKS-QC-8	8	2	10	2	10	0	9	0	10	3	13	4	11	6	11	1	0.82
HUKB-QC-1	10	3	8	0	10	3	8	0	9	3	15	3	9	8	11	0	0.80
HUKB-QC-2	10	6	4	0	9	1	9	0	9	5	15	2	11	8	11	0	0.78
KUIDL-QC-2	4	2	8	1	9	6	5	1	6	5	15	8	8	11	11	0	0.66

8. CONCLUSIONS

NTCIR-10 1CLICK-2 task attracted 10 research teams (including two organizers' teams) from five countries: Japan, China, U.S.A., Canada, and the Netherlands. The total number of English/Japanese Main task submissions is 38, which include 24 English and 14 Japanese runs. The query classification subtask received 21 runs, of which 10 are English and 11 are Japanese. We refer the reader to the 1CLICK-2 participants' papers for details of their runs [9, 11, 12, 14, 16, 17, 25, 26].

9. ACKNOWLEDGMENTS

We thank the NTCIR-10 1CLICK-2 participants for their effort in producing the runs. We would like to thank Lyn Zhu, Zeyong Xu, Yue Dai and Sudong Chung for helping us access to the mobile query log. Part of the English evaluation task was made possibly by NSF Grant IIS-1256172.

10. REFERENCES

[1] J. Aslam. Improving algorithms for boosting. In N. Cesa-Bianchi and S. Goldman, editors, *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, July 2000.

[2] J. A. Aslam, V. Pavlu, and R. Savell. A unified model for metasearch and the efficient evaluation of retrieval systems via the hedge algorithm. SIGIR '03.

[3] Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. *J. Mach. Learn. Res.*, Volume 5, December 2004.

[4] A. Z. Broder. Identifying and filtering near-duplicate documents. COM '00, London, UK, 2000.

[5] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. WWW'97.

[6] B. Carterette, J. Allan, and R. K. Sitaraman. Minimal test collections for retrieval evaluation. SIGIR'06.

[7] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. pages 282–289.

[8] H. Dang and J. Lin. Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors. In *Proc. of ACL 2007*, pages 768–775, 2007.

[9] P. Duboue, J. He, and J.-Y. Nie. Hunter Gatherer: UdeM at 1Click-2. In *Proc. of the 10th NTCIR Conference*, page to appear, 2013.

[10] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. COLT '95.

[11] D. Ionita, N. Tax, and D. Hiemstra. API-based Information Extraction System for NTCIR-1Click. In *Proc. of the 10th NTCIR Conference*, page to appear, 2013.

[12] A. Keyaki, J. Miyazaki, and K. Hatano. XML Element Retrieval@1CLICK-2. In *Proc. of the 10th NTCIR Conference*, page to appear, 2013.

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

[14] T. Manabe, K. Tsukuda, K. Umemoto, Y. Shoji, M. P. Kato, T. Yamamoto, M. Zhao, S. Yoon, H. Ohshima, and K. Tanaka. Information Extraction based Approach for the NTCIR-10 1CLICK-2 Task. In *Proc. of the 10th NTCIR Conference*, page to appear, 2013.

[15] T. Mitamura, H. Shima, T. Sakai, N. Kando, T. Mori, K. Takeda, C.-Y. Lin, R. Song, C.-J. Lin, and C.-W. Lee. Overview of the ntcir-8 aclia tasks: Advanced cross-lingual information access. In *Proc. of NTCIR-8*, pages 15–24, 2010.

[16] H. Morita, R. Sasano, H. Takamura, and M. Okumura. TTOKU Summarization Based Systems at NTCIR-10 1CLICK-2 task. In *Proc. of the 10th NTCIR Conference*, page to appear, 2013.

[17] K. Narita, T. Sakai, Z. Dou, and Y.-I. Song. MSRA at NTCIR-10 1CLICK-2. In *Proc. of the 10th NTCIR Conference*, page to appear, 2013.

[18] V. Pavlu, S. Rajput, P. Golbus, and J. Aslam. Ir system evaluation using nugget-based test collections. In *Proc. of WSDM 2012*, pages 393–402, 2012.

[19] S. Rajput, M. Ekstrand-Abueg, V. Pavlu, and J. A. Aslam. Constructing test collections by inferring document relevance via extracted relevant information. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 145–154, New York, NY, USA, 2012. ACM.

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

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

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

[23] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. *The Annals of Statistics*, 26(5), 1998.

[24] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. COLT '92.

[25] T. Tojima and T. Yukawa. Snippet Summary Similarities Based Query Classification System for the NTCIR-10 1CLICK-2 Task. In *Proc. of the 10th NTCIR Conference*, page to appear, 2013.

[26] M. Yoshioka. Query classification by using named entity recognition systems and clue keywords. In *Proc. of the 10th NTCIR Conference*, page to appear, 2013.

Table 23: Japanese Subtask: Inter-rater agreement in terms of Cohen's kappa coefficient, mean absolute error (MAE), and mean square error (MSE).

assessor pairs		Kappa	MAE	MSE
a_7	a_1	0.782	10.4	2100
a_1	a_2	0.738	15.5	4030
a_2	a_3	0.717	5.32	896
a_3	a_4	0.755	3.87	901
a_4	a_5	0.743	6.01	1480
a_5	a_6	0.751	4.77	744
a_6	a_7	0.688	4.18	1110
a_8	a_9	0.818	4.07	620
a_9	a_{10}	0.770	5.46	599
a_{10}	a_{11}	0.704	6.41	863
a_{11}	a_8	0.753	4.61	731
average		0.747	6.41	1280

APPENDIX

A. QUERIES

Table 21 shows English queries, where overlap queries are indicated by OVERLAP at the column "comment" and specified queries are indicated by "SPEC" at the column "specific". Note that English query IDs are not continuous, as we selected 100 queries from a bigger query pool.

Table 22 shows Japanese queries, where the column "comment" indicates OVERLAP (queries that overlap with English queries), INTENT2 (queries that overlap with NTCIR-10 INTENT2 Japanese queries), INTENT2' (queries that overlap with NTCIR-10 INTENT2 Japanese queries but are slightly changed), or YAHOO (queries that were derived from Yahoo! Chiebukuro[7]).

B. ASSESSMENT RELIABILITY

Table 23 shows inter-rater agreement in terms of Cohen's kappa coefficient, mean absolute error (MAE), and mean square error (MSE) for all the assessor pairs. We hired seven assessors ($a_1 \ldots a_7$) to evaluate submitted runs except MANUAL runs, and four assessors ($a_8 \ldots a_{11}$) for MANUAL runs.

C. QUERIES EXCLUDED IN MANUAL RUNS

As we had generated MANUAL runs (MANUAL-J-D-OPEN-1, MANUAL-J-D-OPEN-2, MANUAL-J-D-OPEN-3, and MANUAL-J-D-OPEN-4) before the Japanese test queries were fixed, the X-string of those runs is empty for the following queries: 1C2-J-0002, 1C2-J-0003, 1C2-J-0004, 1C2-J-0007, 1C2-J-0010, 1C2-J-0012, 1C2-J-0015, 1C2-J-0017, 1C2-J-0018, 1C2-J-0019, 1C2-J-0021, 1C2-J-0024, 1C2-J-0027, 1C2-J-0031, 1C2-J-0032, 1C2-J-0034, 1C2-J-0035, 1C2-J-0036, 1C2-J-0052, 1C2-J-0053, 1C2-J-0054, 1C2-J-0071, 1C2-J-0081, 1C2-J-0083, 1C2-J-0093, 1C2-J-0096, 1C2-J-0097.

[7]Japanese Yahoo! Answer. `http://chiebukuro.yahoo.co.jp/`

Table 21: NTCIR-10 1CLICK-2 English queries.

ID	query type	specific	query	comment
1C2-E-0001	ARTIST	SPEC	michael jackson death	OVERLAP
1C2-E-0002	ARTIST	SPEC	marvin gaye influence	
1C2-E-0004	ARTIST	NONSPEC	dr dre	
1C2-E-0005	ARTIST	NONSPEC	keith sweat	
1C2-E-0006	ARTIST	NONSPEC	glen campbell	
1C2-E-0007	ARTIST	SPEC	whitney houston movies	
1C2-E-0008	ARTIST	NONSPEC	Lil wayne	
1C2-E-0009	ARTIST	NONSPEC	john denver	
1C2-E-0010	ARTIST	NONSPEC	rodney atkins	
1C2-E-0017	ARTIST	NONSPEC	joe arroyo	
1C2-E-0022	ACTOR	NONSPEC	selena gomez	
1C2-E-0025	ACTOR	SPEC	marlon brando acting style	
1C2-E-0026	ACTOR	SPEC	jennifer gardner alias	OVERLAP
1C2-E-0032	ACTOR	SPEC	winston churchill biography	
1C2-E-0033	ACTOR	SPEC	ray charles movie	
1C2-E-0034	ACTOR	SPEC	russell crowe movies	
1C2-E-0035	ACTOR	SPEC	actor the artist	
1C2-E-0036	ACTOR	NONSPEC	charlize theron	
1C2-E-0037	ACTOR	SPEC	keanu reeves family	
1C2-E-0038	ACTOR	SPEC	james cameroon biggest movies	
1C2-E-0042	POLITICIAN	SPEC	robert kennedy cuba	OVERLAP
1C2-E-0045	POLITICIAN	NONSPEC	mayor bloomberg	OVERLAP
1C2-E-0046	POLITICIAN	SPEC	19th president us	
1C2-E-0047	POLITICIAN	NONSPEC	tom corbett	
1C2-E-0048	POLITICIAN	NONSPEC	nancy pelosi	
1C2-E-0049	POLITICIAN	SPEC	ron paul tea party	
1C2-E-0050	POLITICIAN	SPEC	mitt romney governor ma	
1C2-E-0056	POLITICIAN	SPEC	kofi annan syria	
1C2-E-0058	POLITICIAN	SPEC	JFK conspiracy theory	
1C2-E-0061	POLITICIAN	SPEC	hilary clinton first lady	
1C2-E-0062	ATHLETE	NONSPEC	tim tebow	
1C2-E-0063	ATHLETE	SPEC	tom brady records	
1C2-E-0065	ATHLETE	SPEC	aaron rodgers belt celebration	
1C2-E-0070	ATHLETE	NONSPEC	ichiro suzuki	OVERLAP
1C2-E-0071	ATHLETE	SPEC	fabio cannavaro captain	OVERLAP
1C2-E-0073	ATHLETE	SPEC	cristiano ronaldo euro	
1C2-E-0075	ATHLETE	SPEC	lakers coach	
1C2-E-0078	ATHLETE	NONSPEC	tony parker	
1C2-E-0080	ATHLETE	NONSPEC	mario balotelli	
1C2-E-0082	ATHLETE	SPEC	Formula 1 best drivers	
1C2-E-0087	FACILITY	NONSPEC	ut dallas	
1C2-E-0088	FACILITY	NONSPEC	disneyland	
1C2-E-0091	FACILITY	NONSPEC	penn station	
1C2-E-0092	FACILITY	NONSPEC	hawaii pacific university	OVERLAP
1C2-E-0093	FACILITY	NONSPEC	atlanta airport	OVERLAP
1C2-E-0094	FACILITY	SPEC	cheap hotel manhattan july 4	
1C2-E-0095	FACILITY	NONSPEC	american airlines arena	OVERLAP
1C2-E-0096	FACILITY	NONSPEC	TD garden	
1C2-E-0097	FACILITY	SPEC	fedex hub TN	
1C2-E-0098	FACILITY	NONSPEC	french landmarks	
1C2-E-0099	FACILITY	NONSPEC	wimbledon arena	
1C2-E-0100	FACILITY	NONSPEC	wtc twin towers	
1C2-E-0103	FACILITY	SPEC	minnesota bridge collapse	
1C2-E-0105	FACILITY	SPEC	home depot lowes hiring	
1C2-E-0112	FACILITY	NONSPEC	boston duck tour	
1C2-E-0115	GEO	SPEC	theaters texarkana	
1C2-E-0119	GEO	SPEC	apple boylston	
1C2-E-0121	GEO	SPEC	sears illinois	
1C2-E-0125	GEO	SPEC	starbucks san francisco	
1C2-E-0126	GEO	SPEC	bombay christian churches	
1C2-E-0127	GEO	SPEC	Japan earthquake location	
1C2-E-0130	GEO	SPEC	hiphop clubs barcelona	
1C2-E-0131	GEO	SPEC	best summer camping places in US	
1C2-E-0132	GEO	NONSPEC	ski resorts new england	
1C2-E-0133	GEO	SPEC	salsa clubs cali colombia	
1C2-E-0134	GEO	SPEC	concrete delivery nashua NH	
1C2-E-0135	GEO	SPEC	cheap home contractors miami	
1C2-E-0136	GEO	NONSPEC	domino pizza NYC	
1C2-E-0137	GEO	SPEC	best art colleges connecticut	
1C2-E-0140	GEO	NONSPEC	oregon beaches	
1C2-E-0143	DEFINITION	NONSPEC	ewok	
1C2-E-0144	DEFINITION	NONSPEC	geothermal energy	OVERLAP
1C2-E-0145	DEFINITION	NONSPEC	compound interest	
1C2-E-0146	DEFINITION	SPEC	thanksgiving canada	OVERLAP
1C2-E-0148	DEFINITION	NONSPEC	quinoa	
1C2-E-0149	DEFINITION	NONSPEC	sonotubes	
1C2-E-0150	DEFINITION	NONSPEC	cubic yard	OVERLAP
1C2-E-0156	DEFINITION	NONSPEC	enlightenment	
1C2-E-0160	DEFINITION	NONSPEC	batman	
1C2-E-0166	DEFINITION	NONSPEC	rebar	
1C2-E-0167	DEFINITION	NONSPEC	big dig	
1C2-E-0169	DEFINITION	SPEC	sparse matrix	
1C2-E-0170	DEFINITION	NONSPEC	credit score	
1C2-E-0171	DEFINITION	NONSPEC	ip address	
1C2-E-0173	DEFINITION	SPEC	electronic ink	
1C2-E-0178	QA	NONSPEC	why is the sky blue	OVERLAP
1C2-E-0180	QA	SPEC	why do cats purr	OVERLAP
1C2-E-0181	QA	SPEC	why does turkey make you sleepy	
1C2-E-0182	QA	SPEC	why do we yawn	
1C2-E-0183	QA	SPEC	why do flags fly at half mast	
1C2-E-0184	QA	SPEC	why is the ocean salty	OVERLAP
1C2-E-0185	QA	SPEC	where did bloody mary get its name	
1C2-E-0186	QA	SPEC	how does an anteater eat	
1C2-E-0187	QA	SPEC	why are leaves green	
1C2-E-0188	QA	SPEC	what is the difference between weather and climate	
1C2-E-0189	QA	SPEC	why is apple developing maps	
1C2-E-0190	QA	SPEC	what causes sea tide	
1C2-E-0191	QA	SPEC	difference between fission and fusion	
1C2-E-0202	QA	SPEC	why UK does not adopt euro	
1C2-E-0203	QA	NONSPEC	how is trash processed	

Table 22: NTCIR-10 1CLICK-2 Japanese queries.

ID	query type	query	comment
1C2-J-0001	ARTIST	倉木麻衣	INTENT2
1C2-J-0002	ARTIST	ニール・ヤング 来日	INTENT2
1C2-J-0003	ARTIST	太宰治 晩年	
1C2-J-0004	ARTIST	ホイットニー ヒューストン	
1C2-J-0005	ARTIST	小渕 健太郎	
1C2-J-0006	ARTIST	三谷幸喜	
1C2-J-0007	ARTIST	宮部みゆき ドラマ	
1C2-J-0008	ARTIST	高橋 留美子	
1C2-J-0009	ARTIST	梶浦由記	
1C2-J-0010	ARTIST	マイケル ジャクソン 死	OVERLAP
1C2-J-0011	ACTOR	ペネロペクルス	INTENT2
1C2-J-0012	ACTOR	沢尻エリカ 1 リットルの涙	INTENT2
1C2-J-0013	ACTOR	アン ハサウェイ	INTENT2'
1C2-J-0014	ACTOR	市原隼人	
1C2-J-0015	ACTOR	シルヴェスター スタローン	
1C2-J-0016	ACTOR	柳葉敏郎	
1C2-J-0017	ACTOR	栗山千明 カーネーション	
1C2-J-0018	ACTOR	ジェニファー ガーナー エイリアス	OVERLAP
1C2-J-0019	ACTOR	宮崎あおい cm	
1C2-J-0020	ACTOR	沢口靖子	
1C2-J-0021	POLITICIAN	ロバート ケネディ キューバ	OVERLAP
1C2-J-0022	POLITICIAN	亀井静香	
1C2-J-0023	POLITICIAN	谷垣禎一	
1C2-J-0024	POLITICIAN	ブルームバーグ 市長	OVERLAP
1C2-J-0025	POLITICIAN	細野豪志	
1C2-J-0026	POLITICIAN	小沢一郎	
1C2-J-0027	POLITICIAN	小池百合子 キャスター	
1C2-J-0028	POLITICIAN	土井たか子	
1C2-J-0029	POLITICIAN	野田聖子	
1C2-J-0030	POLITICIAN	片山さつき	
1C2-J-0031	ATHLETE	宮里藍 2011 成績	INTENT2
1C2-J-0032	ATHLETE	中村紀洋 ホームラン	INTENT2'
1C2-J-0033	ATHLETE	内田篤人	
1C2-J-0034	ATHLETE	ファビオ カンナヴァーロ キャプテン	OVERLAP
1C2-J-0035	ATHLETE	加藤 康弘 所属クラブ	
1C2-J-0036	ATHLETE	イチロー	OVERLAP
1C2-J-0037	ATHLETE	内村航平	
1C2-J-0038	ATHLETE	岩崎恭子	
1C2-J-0039	ATHLETE	四元奈生美	
1C2-J-0040	ATHLETE	野口みずき	
1C2-J-0041	FACILITY	早稲田大学法科大学院	INTENT2'
1C2-J-0042	FACILITY	京都真如堂	INTENT2'
1C2-J-0043	FACILITY	ホテルアンビア松風閣	INTENT2'
1C2-J-0044	FACILITY	横浜市役所	INTENT2'
1C2-J-0045	FACILITY	南ヶ丘牧場	INTENT2'
1C2-J-0046	FACILITY	須磨海浜水族園	INTENT2'
1C2-J-0047	FACILITY	未来科学館	
1C2-J-0048	FACILITY	カーサ・ディ・ナポリ	
1C2-J-0049	FACILITY	ザ・ペニンシュラ東京	
1C2-J-0050	FACILITY	小金井図書館	
1C2-J-0051	FACILITY	あおやま矯正歯科医院	
1C2-J-0052	FACILITY	アメリカン エアラインズ アリーナ	OVERLAP
1C2-J-0053	FACILITY	アトランタ 空港	OVERLAP
1C2-J-0054	FACILITY	ハワイパシフィック大学	OVERLAP
1C2-J-0055	FACILITY	らーめんてつや	
1C2-J-0056	GEO	宇都宮駅 焼き鳥	
1C2-J-0057	GEO	大宮駅周辺 天ぷら	
1C2-J-0058	GEO	松山市 イタリアン	
1C2-J-0059	GEO	映画館 名古屋	
1C2-J-0060	GEO	岡山駅 カラオケ	
1C2-J-0061	GEO	京都市 スーパー銭湯	
1C2-J-0062	GEO	明石市 整形外科	
1C2-J-0063	GEO	東淀川区 眼科	
1C2-J-0064	GEO	静岡市 駅前 産婦人科	
1C2-J-0065	GEO	江ノ島 旅館	
1C2-J-0066	GEO	博多駅 ホテル	
1C2-J-0067	GEO	千葉駅 郵便局	
1C2-J-0068	GEO	川口市 交番	
1C2-J-0069	GEO	札幌 美容専門学校	
1C2-J-0070	GEO	恵比寿 結婚式場	
1C2-J-0071	DEFINITION	地熱発電	INTENT2,OVERLAP
1C2-J-0072	DEFINITION	犬夜叉	INTENT2
1C2-J-0073	DEFINITION	gps	INTENT2
1C2-J-0074	DEFINITION	秘密の花園	INTENT2
1C2-J-0075	DEFINITION	トルネード	INTENT2
1C2-J-0076	DEFINITION	avp	INTENT2
1C2-J-0077	DEFINITION	ザ・ウォール	INTENT2
1C2-J-0078	DEFINITION	バズライトイヤー	INTENT2
1C2-J-0079	DEFINITION	急がば回れ	INTENT2'
1C2-J-0080	DEFINITION	クフ王	INTENT2
1C2-J-0081	DEFINITION	感謝祭 カナダ	OVERLAP
1C2-J-0082	DEFINITION	パーキンソン病	INTENT2'
1C2-J-0083	DEFINITION	立方ヤード	INTENT2,OVERLAP
1C2-J-0084	DEFINITION	国際会計基準	INTENT2'
1C2-J-0085	DEFINITION	光ファイバー	INTENT2'
1C2-J-0086	QA	プラズマと液晶の違い	INTENT2
1C2-J-0087	QA	もち米の炊き方	INTENT2
1C2-J-0088	QA	円柱の体積を求める公式	INTENT2
1C2-J-0089	QA	ハヤシライスの作り方	INTENT2
1C2-J-0090	QA	寒中見舞いの文例	INTENT2
1C2-J-0091	QA	牡蠣料理のレシピ	INTENT2
1C2-J-0092	QA	長ネギの青い部分に詰まっている透明なゼリー状の物質は何か	YAHOO
1C2-J-0093	QA	なぜ空は青いのか	OVERLAP
1C2-J-0094	QA	エコノミークラス症候群予防方法	YAHOO
1C2-J-0095	QA	四葉のクローバーが幸運を呼ぶ御守りである理由	YAHOO
1C2-J-0096	QA	なぜ海はしょっぱいのか	OVERLAP
1C2-J-0097	QA	なぜ猫はのどを鳴らすのか	OVERLAP
1C2-J-0098	QA	世界で初めてノーベル賞をとった人は誰か	YAHOO
1C2-J-0099	QA	「くすぐったい」という感覚はどのようにして引き起こされるか	YAHOO
1C2-J-0100	QA	盛り塩をする意味	YAHOO