Overview of NTCIR-9 1CLICK

Tetsuya Sakai† Makoto P. Kato* Young-In Song†
†Microsoft Research Asia *Kyoto University
tetsuyasakai@acm.org, kato@dl.kuis.kyoto-u.ac.jp

Abstract

This is an overview of the NTCIR-9 One Click Access (1CLICK) "pilot" task. 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. Systems are expected to present important pieces of information first and to minimise the amount of text the user has to read. As our first trial, we designed a Japanese 1CLICK task with a nugget-based test collection. Three research teams participated in the task, using diverse approaches (information extraction, passage retrieval and summarisation). Our results suggest that the 1CLICK evaluation framework is a useful complement to traditional 10-blue-link evaluation. We therefore hope to expand the language scope in the next round, at NTCIR-10.
Keywords: *test collections, information access, mobile, desktop, nuggets, evaluation metrics, S-measure.*

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. Systems are expected to present important pieces of information first and to minimise the amount of text the user has to read.

Our recent CIKM paper discusses the motivation behind the design of the 1CLICK task, how the NTCIR-9 1CLICK test collection was constructed, as well how the evaluation methods were devised [5]. We recommend the reader to read the CIKM paper first, as the present overview serves as a complement to that paper, by presenting the outcome of the official results for the 1CLICK participants.

Table 1 provides a list of the NTCIR-9 1CLICK participants. Unfortunately, we only had three participating teams (even though 25 teams signed up for 1CLICK!), from which we received a total of 10 runs. As our evaluation framework is basically language-independent, we hope to expand our language scope at NTCIR-10.

NTCIR-9 Workshop Meeting, 2011, Tokyo, Japan.

Table 1. 1CLICK participants.

team name	organisation
KUIDL	Kyoto University
MSRA1click	Microsoft Research Asia
TTOKU	Tokyo Institute of Technology

The important dates for NTCIR-9 1CLICK were as follows:

March 1	*Training queries (40) and nuggets released*
April 28	*Formal run queries (60) released*
May 16	*Run submissions due*
May 23	*Formal run nuggets (tentative version) released*
June 13	*Feedback on formal run nuggets due*
July 1	*Nugget match evaluation begins*
August 31	*Nugget match evaluation ends*
September 8	*Formal run results released*

The "rebuttal" period (May 23 – June 13) was possibly unsuccessful: the organisers sent the gold-standard nuggets to participants, and the participants were asked for feedback. However, as the organisers did not receive any feedback, the "tentative" nuggets became the official nuggests without any modifications whatsoever. Also, the 1CLICK evaluation framework requires assessors to conduct *nugget match evaluation*: the process of manually comparing a system output and a list of gold-standard nuggets by means of a dedicated interface (See Section 4.1). The organisers and the participants both worked as assessors between July 1 and August 31, although the actual total workload per assessor was only 302 minutes on average. Finally, using the results of nugget match evaluation, a new evaluation metric called *S-measure* was computed along with *weighted recall* [5].

The remainder of this paper is organised as follows. Section 2 describes the specifications of the NTCIR-9 1CLICK task, and Section 3 describes the NTCIR-9 1CLICK Japanese test collection. Section 4 describes our evaluation framework, and Section 5 reports on the official results of 1CLICK. Finally, Section 6 concludes this paper and discusses future work.

2. Task

As mentioned earlier, 1CLICK systems should aim to satisfy the user with a single textual output, immediately after the user clicks on the SEARCH button. They are expected to present important pieces of information first and to minimise the amount of text the user has to read. This section describes the formats of the input to and the output from the systems. As this round of 1CLICK concerns Japanese texts, we used UTF-8 as the encoding scheme

throughout the task.

2.1 Input

The input to a 1CLICK system is a query file in which each line is of the following form.

```
<queryID> <querystring>
```

There are four query types, and for each type we assume that the user has the following information needs:

CELEBRITY User wants to gather various facts about a celebrity: date/place of birth, real name, blood type, height, hobbies, profession, personal history, awards, publications, discography, films, TV series, favourite baseball team, favourite food etc.

LOCAL User wants to contact or visit a facility (school, shop, office, amusement park, hotel, train station etc.). Hence (s)he wants facts such as postal and email addresses, phone and fax numbers, opening hours, how to access the facility by train/bus/car, nearest stations, time required for the travel, whether the facility has a car park and its opening hours etc.

DEFINITION User seeks the definition of a term.

QA User seeks a short answer to a question.

These query types were chosen based on a previous study on Japanese mobile and desktop query logs [2].

For each query, the system is expected to gather information from the web or any other knowledge sources. Runs produced by following this approach are called the *Open* runs.

Each query is associated with a set of gold-standard nuggets, and each nugget has a supporting URL [5]. Therefore, participating systems can optionally treat these URLs as an additional set of input to produce an output. Runs produced by following this approach are called the *Oracle* runs. Note that oracle runs treat both the query and its "right answer" URLs as input, just like query-oriented multi-document summarisation.

The test query file contained 60 queries (15 CELEBRITY, 15 LOCAL, 15 DEFINITION and 15 QA). In addition, prior to the formal run evaluation, the organisers released 40 queries (10 for each query type) with nuggets to participants[1].

2.2 Output

For each input query, 1CLICK systems are expected to produce a plain text of X characters, excluding punctuation marks, special symbols etc. We call this the X-string. We allowed two types of output: DESKTOP runs ("D-runs") where X is set to 500; and MOBILE runs ("M-runs") where X is set to 140. The former roughly corresponds to five Japanese search engine snippets which the user can typically view without scrolling the browser window (i.e., those "above the fold."); the latter approximates a mobile phone display size.

Thus four types of run were possible: Open D-runs, Open M-runs, Oracle D-runs and Oracle M-runs. The run file name was required to be of the following form:

```
<teamID>-<runtype>-<source>-<integer>.txt
```

where

`runtype` was either D (DESKTOP) or M (MOBILE); and

`source` was either OPEN or ORCL (Oracle).

Manual runs were not allowed. The exact run file format is described in the 1CLICK homepage[2].

3. Test Collection

The NTCIR-9 1CLICK test collection consists of 60 test queries[3] shown in Figure 1, and a set of gold-standard nuggets for each query. The CELEBRITY and LOCAL queries were selected from a proprietary Japanese mobile query log, while the DEFINITION and QA queries were devised based on Yahoo! Chiebukuro (Japanese Yahoo! answers) data [4]. A separate set of 40 "training" queries were created in a similar way.

A 1CLICK nugget is a 5-tuple:

```
<nuggetID> <weight> <semantics>
                <vital string> <URL>
```

where

`weight` is the importance of the nugget. Two Japanese 1CLICK organisers and three Japanese participants independently weighted each nugget using a 3-point scale (1-3) and the sum was used for evaluation[4].

`semantics` is the factual statement that the nugget conveys. This is used by the assessor to determine whether a nugget is present in the X-string or not.

`vital string` is a short piece of text that is probably necessary to be included in the X-string in order to convey the meaning of the nugget. This is used for defining a *Pseudo Minimal Output* [5], which approximates a most consise output that covers all nuggets and orders them appropriately. The PMO is used for normalising our evaluation metric.

`URL` is a "supporting document" for the nugget. Note that Oracle runs treat these URLs as a part of input to the system.

Figure 2 shows some actual nuggets from one of our test queries. Nuggets N001 and N003 say that Osamu Tezuka was born on November 3, 1928 and died on February 9, 1989, respectively; N002 says that he was born in Osaka; N004 says that he was a cartoonist; N009 and N013 say that he graduated from Osaka University in 1951 and that he got married in 1959; N014 says that his wife's name is Etsuko; and N015 says that he received a medical doctoral degree in 1961. The English translations of the corresponding vital strings would be: "Nov 3, 1928," "Osaka," "Feb 9,

[1]The first author of this paper created the 60 test queries and their nuggets; a vendor created the 40 training queries and their nuggets by using the test queries for reference and following the first author's instructions.

[2]`http://research.microsoft.com/en-us/people/tesakai/1click.aspx`

[3]Query 0036 is misspelt, but we did not correct it. Current Web Search APIs can actually correct the spelling automatically.

[4]Our original intention was to use the two sets of weights from the organisers, but it turned out that three participants unexpectedly completed their own nugget weight assignments on the nugget match interface we released (See Section 4.1). We really thank their diligence!

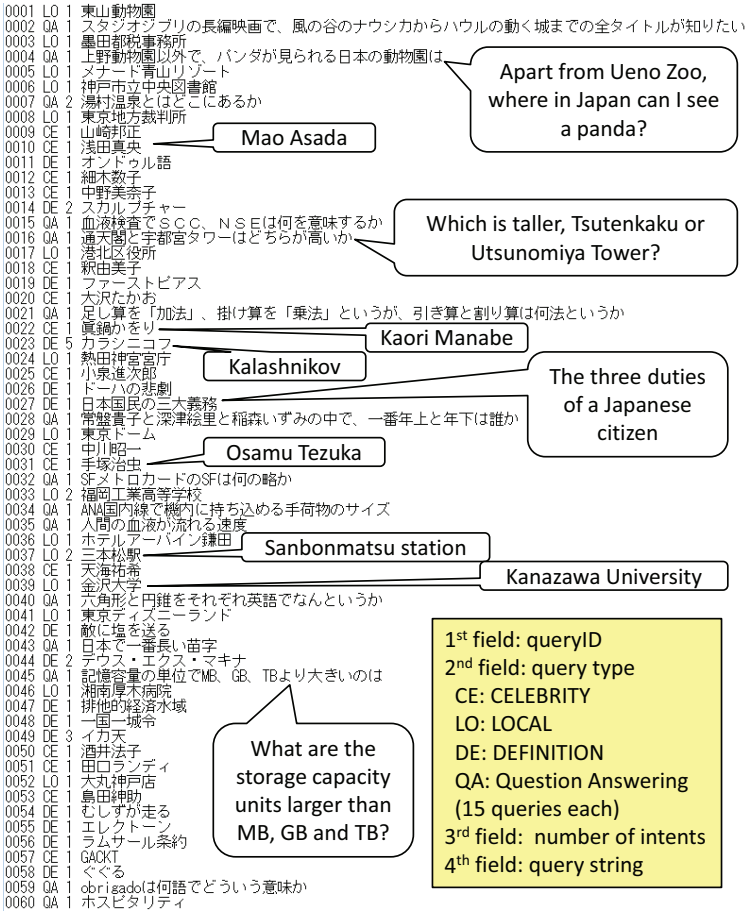

Figure 1. NTCIR-9 1CLICK test queries.

N001	6	1928年11月3日生	1928.11.03	http://tezukaosamu.net/jp/about/index.html
N002	6	出身地 大阪府	大阪	http://tezukaosamu.net/jp/about/index.html
N003	6	1989年2月9日没	1989.02.09	http://tezukaosamu.net/jp/about/index.html
N004	6	漫画家	漫画家	http://tezukaosamu.net/jp/about/index.html
N009	5	1951年 大阪大学卒業	卒業	http://tezukaosamu.net/jp/about/1950.html
N013	3	1959年 結婚	結婚	http://tezukaosamu.net/jp/about/1950.html
N014	2	妻 悦子	悦子	http://tezukaosamu.net/jp/about/1950.html
N015	5	1961年 医学博士号取得	医学博士	http://tezukaosamu.net/jp/about/1960.html

Figure 2. Some nuggets from Query 0031 "Osamu Tezuka" (CELEBRITY).

1989," "cartoonist," "graduated," "married," "Etsuko" and "medical doctor."

All nuggets were designed to represent facts as of December 31, 2010. We refer the reader to our CIKM paper [5] for more details.

4. Evaluation Methods

The 1CLICK evaluation requires two steps. The first is to use the nugget match interface for manually determining which nuggests are included in the X-string. This process also records the position of each nugget found in the X-string. The second step is to compute evaluation metrics based on the gold standard nuggets and the result of nugget matching.

4.1 Nugget Match Interface

Figure 3 shows the nugget match evaluation interface we developed for 1CLICK[5]. The primary purpose of this interface is to let the assessor compare the X-string with the list of gold-standard nuggets and determine the presence and position of each nugget. In the figure, the nugget that represents the fact that "Keiko Kitagawa's date of birth is August 22, 1986" has been identified within the X-string and the position information has been recorded: "[184, 199]" represent the start and end positions of the highlighted area (called the *nugget match area*).

The nugget position information is later used for computing our primary evaluation metric, *S-measure*. However, as S-measure merely evaluates the nugget ranking within the X-string, our interface also provides radio buttons for assessing the *readability* and the *trustworthiness* of each X-string, as shown at the top of Figure 3:

Readability is to do with coherence and cohesiveness, and how easy it is for the user to read and understand the text. For

[5]We have made several changes compared to our first interface [5].

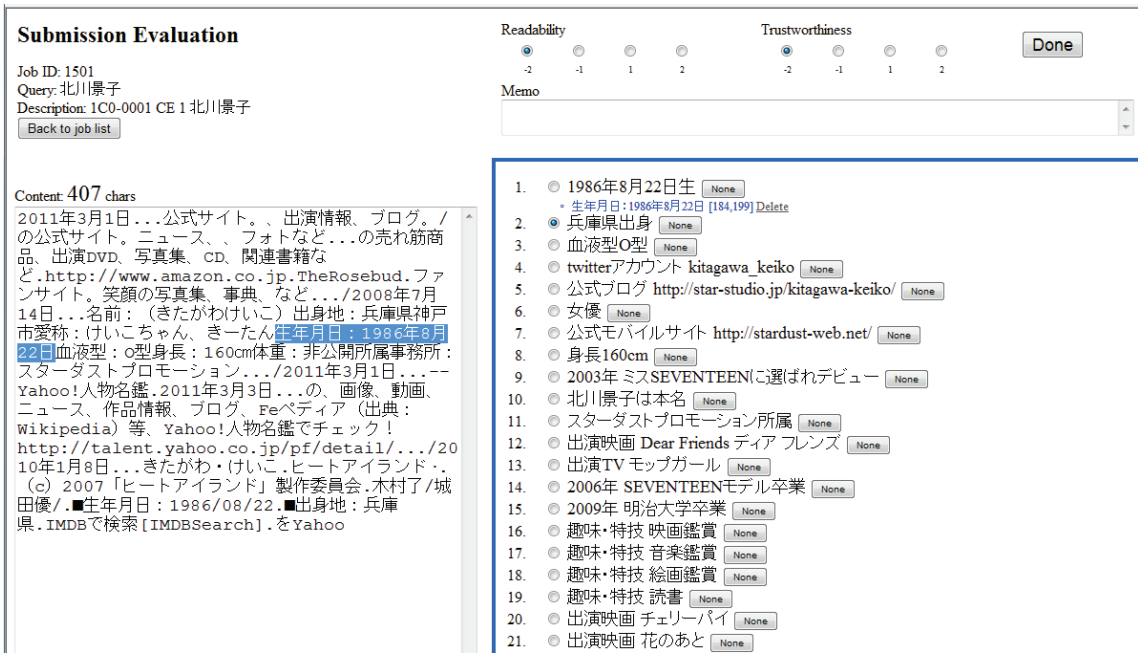

Figure 3. Nugget match interface.

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.

In principle, relevance, readability and trustworthiness should be orthogonal to one another. Detailed instructions for the nugget match assessors can be found in the 1CLICK homepage[6].

We had 60 queries × 10 runs and therefore 600 X-strings to evaluate. We had 10 assessors (organisers and participants who are fluent in Japanese), and each assessor evaluated 120 randomised X-strings: that is, each X-string was independently assessed by two assessors.

Our interface recorded, for each assessor, the total time he spent on assessing each X-string. The interface allowed assessors to abort and restart their jobs at any time.

4.2 S-measure

The output from the nugget match interface is a set of *matched nuggets* and their position information. Based on this, we compute S-measure as follows.

Let N be the set of gold-standard nuggets constructed for a particular query, and let $v(n)$ be the vital string and let $w(n)$ be the weight for nugget $n \in N$. The Pseudo Minimal Output (PMO) for this query is defined by sorting all vital strings by $w(n)$ (first key) and $|v(n)|$ (second key) [5]. Let $\mathit{offset}^*(v(n))$ denote the offset position of $v(n)$ within the PMO.

Let $M(\subseteq N)$ denote the set of *matched* nuggets, and let $\mathit{offset}(m)$ denote the offset position of $m \in M$. Morever, let L be a parameter that represents how the readers's patience runs out: we let $L = 500$ in this paper, which means that the X-string is of no value after 500 characters. Since it is known that the average reading speed of a Japanese person is around 400-600 characters per minute, this setting means that the user has about one minute to examine the X-string, but no more.

S-measure can then be computed as follows[7]:

$$\mathit{S\text{-}measure} = \frac{\sum_{m \in M} w(m) \max(0,\ L - \mathit{offset}(m))}{\sum_{n \in N} w(n) \max(0,\ L - \mathit{offset}^*(v(n)))} . \quad (1)$$

We can also compute *weighted recall* by removing the "max factors" from the above equation. Note that while S-measure is a ranked retrieval metric for nuggets, weighted recall is a set retrieval metric for nuggets.

Figure 4 provides an example of how S-measure is computed. (This is an example deliberately chosen to highlight possible limitations with S-measure [5].) This query has only four nuggets, two weighted 6 and two weighted 4. The PMO is defined by arranging the four nuggets as <N003, N001, N004, N002> by sorting them first by the weight and then by the vital string length. The score for this PMO is computed as shown in the figure, and amounts to

[6] http://research.microsoft.com/en-us/people/tesakai/nuggetmatch-v1.txt

[7] Our CIKM paper [5] also describes "S-flat," defined as $\max(1, \mathit{S\text{-}measure})$, as S-measure is not theoretically bounded above by 1. However, for all of our official results, S-measure values were below one and therefore the "flattening" was not necessary.

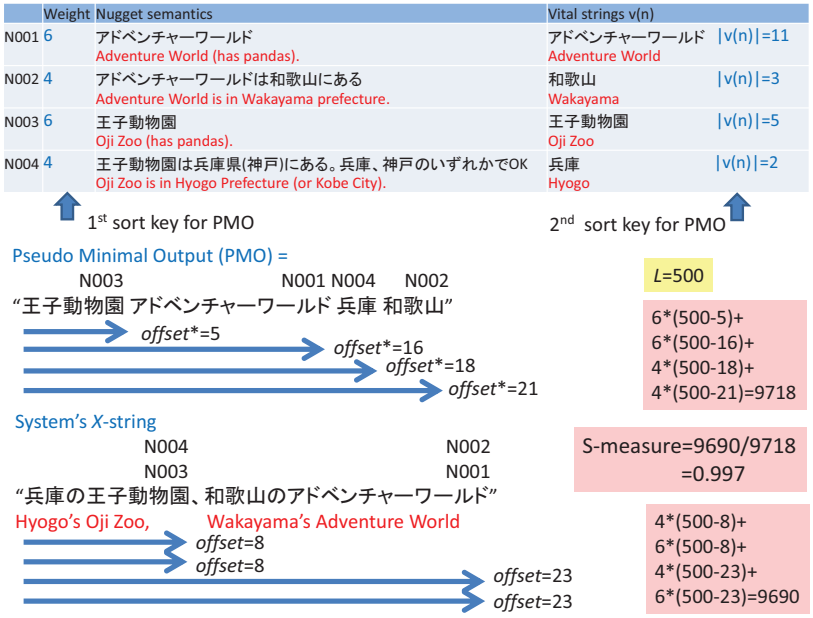

Figure 4. Computing S-measure: a somewhat problematic example (Query 0004 "Apart from Ueno Zoo, where in Japan can I see a panda? [English translation]").

9718. Now, suppose that we have an X-string as shown in the same figure. This string covers all four nuggets and therefore its weighted recall is exactly one. Whereas, its S-measure value is a little smaller than one, as the nugget positions are taken into account as follows. Suppose that the offsets for nuggets N003 and N004 are defined (by the nugget match assessor) as 8. (Note that the offset for N004 (Hyogo) is not 2: the string "Hyogo" alone does not convey the meaning of N004 "Oji Zoo is in Hyogo Prefecture.") Similarly, suppose that the offsets for nuggets N002 and N001 are defined as 23 (counted after excluding the Japanese comma). The unnormalised score for this X-string is computed as shown in the figure, and its S-measure is $9690/9718 = 0.997$.

While the advantages of using S-measure are not altogether clear from the above particular example, S-measure can penalise redundant text near the beginning of the X-string, and also reward systems that present important nuggets first, unlike weighted recall [5]. We shall come back to this when we discuss the official results in Section 5.2.

As each X-string was independently assessed by two assessors, we computed the following four versions of S-measure:

A S-measure computed based on nugget matches identified by "Assessor A";

B S-measure computed based on nugget matches identified by "Assessor B";

I S-measure computed based on the intersection between the above two sets of nugget matches. For each nugget match, the offset is defined as the average of **A**'s offset and **B**'s offset.

U S-measure computed based on the union of the nugget matches by **A** and **B**. For each nugget match identified by both **A** and **B**, the offset is defined as in **I**; for each nugget match identified by only one assessor, the offset is taken as is.

Similarly, we computed four versions of weighted recall (W-recall) as well. Note that "Assessors" A and B are actually different people across queries due to randomisation (See Section 4.1).

5. Results

5.1 Submitted Runs

Table 2 shows the names of the 10 runs submitted, together with their SYSDESC (system description) fields. Details of the individual runs are described in the 1CLICK participants's papers [6, 7, 8]. Note that the four runs from KUIDL and the two runs from MSRA1click are Open runs, while the four runs from TTOKU are Oracle runs. Thus, TTOKU treated the URL lists of nuggets as part of input to their system.

5.2 Official Results

Table 3 shows the official mean S-measure performances of the submitted runs. The runs are ranked by the mean S-measure with **I**. Similarly, Table 4 shows the mean W-recall performances. Figures 5 and 6 visualise these two tables. It can be observed that, on average, KUIDL-D-OPEN-1 is the overall winner both in terms of S-measure and W-recall. Moreover, the trends across the four nugget match data **I**, **U**, **A** and **B** are relatively similar.

The set of 10 runs were tested for statistical significance with S-measure and W-recall (with I and U) using a randomised version of two-sided Tukey's Honestly Significant Differences (HSD)

Table 2. 10 1CLICK runs submitted. (The SYSDESCs for KUIDL turned out to be identical.)

run name	SYSDESC field (line 1 of the run file)
KUIDL-D-OPEN-1	SVM based query classification; various Yahoo! APIs search results; LexRank based ranking; MMR based selection
KUIDL-D-OPEN-2	SVM based query classification; various Yahoo! APIs search results; LexRank based ranking; MMR based selection
KUIDL-M-OPEN-1	SVM based query classification; various Yahoo! APIs search results; LexRank based ranking; MMR based selection
KUIDL-M-OPEN-2	SVM based query classification; various Yahoo! APIs search results; LexRank based ranking; MMR based selection
MSRA1click-D-OPEN-1	query classification → (BingAPI with query expansion→ segment merging | Wikipedia | YahooChiebukuro)
MSRA1click-D-OPEN-2	query classification → (BingAPI without query expansion→ segment merging | Wikipedia | YahooChiebukuro)
TTOKU-D-ORCL-1	This system makes an ILP problems to make an abstractive summary.
TTOKU-D-ORCL-2	Use Max-min problem to cover informations about query type
TTOKU-M-ORCL-1	This system makes an ILP problems to make an abstractive summary.
TTOKU-M-ORCL-2	Use Max-min problem to cover informations about query type

test [1]. Figures 7 and 8 list up the significantly different pairs at $\alpha = 0.05$. Of the 45 run pairs, 20 pairs are significantly different in terms of S-measure with **I**; 19 pairs are significantly different in terms of W-recall with **I**; 18 pairs are significantly different in terms of S-measure with **U**; 20 pairs are significantly different in terms of W-recall with **U**. Thus the discriminative power is similar for all four metrics, but it can be observed in the figures that sometimes S and W disagree as to which run pairs are significant. It can be observed that none of the differences between the two teams KUIDL and MSRA1click is statistically significant.

Table 5 shows the mean performances by query type (CE, LO, DE and QA) with **I** and **U**. Figures 9-12 visualise these results. Note that the performances with **U** are by definition always higher than the corresponding performances with **I**, as **U** implies "more nugget matches than **I**."

It can be observed that the strength of the top performer KUIDL-D-OPEN-1 comes from its ability to handle CE and QA queries. Its performance for the CE queries are particularly impressive. On the other hand, it can be observed that MSRA1click-D-OPEN-2 does very well with LO and DE queries. In addition, it can be observed that the QA queries are generally easy, while the LO queries are generally hard. The performances for the QA queries are generally high probably because the participating teams rely on community QA data (Yahoo! Chiebukuro) for these queries. Recall that the queries (but not necessarily the nuggets) actually originate from the Yahoo! Chiebukuro data.

It can also be observed that, when averaged per query type, the ranking by S-measure and that by W-recall can be quite different. For example, in Table 5(c) and Figure 11, while TTOKU-D-ORCL-1 is ranked at eight in terms of mean S-measure with **I** for the DE queries, it is actually the third best run for DE in terms of W-recall with **U**. This implies that while this run managed to cover many nuggets, it did not order them appropriately. We will discuss per-query differences between S-measure and W-recall later in Section 5.4.

5.3 Per-query Inter-assessor Disagreements

The Cohen's Kappa values between two different assessors (recall that we had ten assessors) were generally high: they varied between 0.68 and 0.88.

To closely examine per-query inter-assessor disagreements on nugget matches, we selected the "best" run from each team in terms of mean S-measure with **I**. Figure 13 visualises the per-query inter-assessor disagreements in terms of S-measure.

Figure 14 provides a detailed diagnosis for each baloon shown in Figure 13, i.e., cases where there were substantial inter-assessor disagreements. Nuggets identified by Assessors A and B are shown in blue, and the nugget match areas specified by either of the assessors are shown in red within the actual X-string. Our comments are shown in baloons. Below, we discuss all of these cases by referring to Figures 14 and 13.

For KUIDL-D-OPEN-1 / Query 0007, it is clear that Assessor A missed an existing nugget and that is why the S-measure with **A** is zero. In contrast, for Query 0059, Assessor B, who didn't find any nuggets in the X-string, is probably right, and we regard Assessor A's Nugget N002 is a false alarm. N002 represents the fact "obrigado means thank you" but the X-string does not actually convery this. It abruptly starts with "Thank you, master." Thus we believe that the S-measure should really be zero for this case. Similarly, for MSRA1click-D-OPEN-2, Assessor A found Nugget N003 but this is a false alarm. These examples demonstrate that hiring multiple assessors is worthwhile.

A more subtle case is MSRA1click-D-OPEN-2 / Query 0054. This is a DE query, so the X-string is supposed to provide the meaning of the Japanese idiom. Assessor A found N003 which represents the fact that *Mushizu ga hashiru* can mean "to have heartburn." However, we argue that this is probably a false alarm, because the X-string only says that *Mushizu* means regurgitated gastric acid, and does not explain what the entire idiom means. There may be room for improvement in the nugget match assessment guideline as well as in how we formulate the nugget semantics.

Another interesting example is TTOKU-D-ORCL-1 / Query 0019. While Assessor A found three nuggets (correctly, in our view), Assessor B found none and therefore the S-measure with **B** is zero. We conjecture that this is because the readability of the X-string is quite poor, as this run used an abstractive summarisation approach which often generated ungrammatical sentences. Note that a real-world user may be more like Assessor B: she will probably not bother to read a barely readable text unless she is desperate. We will discuss readability and trustworthiness in Section 5.5.

Table 3. Mean S-measure over 60 queries: runs ranked by I. The highest value within each column is shown in bold.

	I	U	A	B
KUIDL-D-OPEN-1	**0.3132**	**0.3814**	**0.3597**	**0.3347**
MSRA1click-D-OPEN-2	0.2988	0.3268	0.3186	0.3069
KUIDL-D-OPEN-2	0.2900	0.3467	0.3166	0.3199
MSRA1click-D-OPEN-1	0.2832	0.3285	0.3041	0.3075
KUIDL-M-OPEN-2	0.2214	0.2730	0.2420	0.2524
KUIDL-M-OPEN-1	0.2196	0.2834	0.2467	0.2563
TTOKU-D-ORCL-1	0.1585	0.1969	0.1851	0.1702
TTOKU-D-ORCL-2	0.1484	0.2316	0.2136	0.1662
TTOKU-M-ORCL-1	0.0866	0.1418	0.1168	0.1116
TTOKU-M-ORCL-2	0.0829	0.1312	0.1148	0.0993

Table 4. Mean W-recall over 60 queries: runs ranked by I. The highest value within each column is shown in bold.

	I	U	A	B
KUIDL-D-OPEN-1	**0.3468**	**0.4236**	**0.3970**	0.3734
KUIDL-D-OPEN-2	0.3413	0.4074	0.3741	**0.3747**
MSRA1click-D-OPEN-2	0.3088	0.3391	0.3305	0.3174
MSRA1click-D-OPEN-1	0.2826	0.3359	0.3091	0.3094
TTOKU-D-ORCL-1	0.2321	0.2851	0.2663	0.2510
KUIDL-M-OPEN-2	0.2147	0.2624	0.2307	0.2463
KUIDL-M-OPEN-1	0.2043	0.2646	0.2286	0.2403
TTOKU-D-ORCL-2	0.1704	0.2610	0.2392	0.1922
TTOKU-M-ORCL-1	0.0921	0.1493	0.1224	0.1190
TTOKU-M-ORCL-2	0.0779	0.1211	0.1087	0.0903

Our final example is TTOKU-D-ORCL-1 / Query 0028, where Assessor A clearly missed a nugget (N001), and as a result, also missed N005 (in our view). This is a QA question that asks who is the oldest/youngest among the three Japanese actresses. According to our policy ("*If any person with common sense can clearly and immediately judge that the nugget is true by reading the context in the X-string, this may be counted as a nugget match.*), if the X-string contains the birthdays of all three actresses, then it answers the questions. Nevertheless, "common sense" is a grey area in general.

To sum up our inter-assessor disagreement analysis: (a) Hiring multiple assessors is useful as nugget misses and false alarms can sometimes happen; and (b) The nugget match criteria and the formulation of nugget match semantics may deserve further elaborations.

5.4 Per-query Inter-metric Disagreements

To closely examine the differences between S-measure and W-recall, Figure 15 shows the per-query S-measure and W-recall values based on **I** for KUIDL-D-OPEN-1 and TTOKU-D-ORCL-1. (We also did a similar analysis for MSRA1click-D-OPEN-2 but the results were less interesting as the differences between S-measure and W-recall values were smaller.)

Figure 16 provides a diagnosis for each baloon shown in Figure 15, i.e., cases where there were substantial differences between S-measure and W-recall values with **I**. Nugget identified by both assessors are shown in blue and the corresponding nugget match areas are shown in red. It can be observed that the reason why S-measure is much lower than W-recall for these cases is that there is a lot of nonrelevant text in the X-strings before relevant nuggets appear. In particular, the firt part of the X-string for the first example (KUIDL / 0019) is completely off-topic; and the relevant information appears at the very end of the X-string for the second and the third examples. These examples demonstrate that evaluation with S-measure is useful for carefully designing textual output strategies: this may apply not only to 1CLICK systems but also to snippet generation, hover preview generation, and query-oriented summarisation.

5.5 Readability and Trustworthiness

Table 6 shows the mean readability and trustworthiness values for the ten runs. The mean is taken across queries and across assessors. (For TTOKU-D-ORCL-1, Query 0033 was omitted as the run had a formatting error for this query.) Recall that the raw readability and trustworthiness values range from -2 to 2 (See Figure 3). The mean S-measure values based on **I** are also shown again for comparison. It is interesting that the mean readability for KUIDL-M-OPEN-2 is higher than KUIDL-D-OPEN-2 and in fact higher than any other runs: note, for example, that a single short sentence may be perfectly readable regardless of its relevance.

Figure 17 visualises Table 6, but focusses on the six D-runs for clarity. Note that while mean S-measure values are always positive, mean readability and trustworthiness can be negative. It can be observed that the mean readability of TTOKU-D-ORCL-1 is very low compared to the other runs. This is probably because TTOKU explored abstractive approaches while others used extractive approaches [8]. Figures 19-20 show the complete list of comments we obtained through the nugget match evaluation phase: the comments for TTOKU-D-ORCL-1 indeed suggest that its X-strings often contained incomplete sentences. Although TTOKU-D-ORCL-1 does not appear to be successful for this year, we believe that research in abstractive approaches are very important for

Table 5. Mean S-measure/W-recall by query type: runs ranked by S-measure (I). The highest value within each column is shown in bold.

	S-measure (**I**)	W-recall (**I**)	S-measure (**U**)	W-recall (**U**)
(a) 15 CE queries				
KUIDL-D-OPEN-1	**0.2269**	**0.1591**	**0.2739**	**0.1919**
MSRA1click-D-OPEN-2	0.1523	0.0985	0.1885	0.1357
MSRA1click-D-OPEN-1	0.1431	0.0967	0.1667	0.1219
KUIDL-D-OPEN-2	0.1383	0.1039	0.1840	0.1381
KUIDL-M-OPEN-1	0.1363	0.0732	0.1837	0.1027
TTOKU-D-ORCL-2	0.1276	0.0881	0.2100	0.1414
KUIDL-M-OPEN-2	0.0903	0.0520	0.1185	0.0701
TTOKU-M-ORCL-2	0.0675	0.0371	0.1198	0.0620
TTOKU-D-ORCL-1	0.0215	0.0156	0.0401	0.0297
TTOKU-M-ORCL-1	0.0071	0.0053	0.0215	0.0153
(b) 15 LO queries				
MSRA1click-D-OPEN-2	**0.1678**	**0.2081**	0.2079	0.2535
MSRA1click-D-OPEN-1	0.1674	0.1596	0.2052	0.2137
KUIDL-D-OPEN-2	0.1607	0.1941	0.2578	0.3113
KUIDL-D-OPEN-1	0.1513	0.1777	**0.2619**	**0.2920**
TTOKU-D-ORCL-2	0.1473	0.1752	0.1749	0.2094
TTOKU-D-ORCL-1	0.1207	0.1957	0.1623	0.2409
TTOKU-M-ORCL-2	0.0972	0.0905	0.1165	0.1033
TTOKU-M-ORCL-1	0.0770	0.0720	0.1126	0.1029
KUIDL-M-OPEN-1	0.0765	0.0611	0.1132	0.0878
KUIDL-M-OPEN-2	0.0689	0.0551	0.1357	0.1079
(c) 15 DE queries				
MSRA1click-D-OPEN-2	**0.3353**	0.3700	**0.3709**	0.4088
KUIDL-D-OPEN-1	0.3216	**0.4061**	0.3544	0.4604
KUIDL-D-OPEN-2	0.2911	0.4047	0.3358	**0.4605**
MSRA1click-D-OPEN-1	0.2795	0.3157	0.3669	0.4128
KUIDL-M-OPEN-2	0.2597	0.2754	0.2992	0.3186
TTOKU-D-ORCL-2	0.2208	0.2905	0.2927	0.4050
KUIDL-M-OPEN-1	0.2109	0.2184	0.2981	0.3149
TTOKU-D-ORCL-1	0.2005	0.3266	0.2728	0.4489
TTOKU-M-ORCL-1	0.0825	0.0952	0.2013	0.2259
TTOKU-M-ORCL-2	0.0556	0.0588	0.0941	0.1049
(d) 15 QA queries				
KUIDL-D-OPEN-2	**0.5698**	**0.6626**	0.6094	0.7198
KUIDL-D-OPEN-1	0.5529	0.6445	**0.6355**	**0.7503**
MSRA1click-D-OPEN-1	0.5429	0.5585	0.5754	0.5955
MSRA1click-D-OPEN-2	0.5399	0.5585	0.5399	0.5585
KUIDL-M-OPEN-2	0.4667	0.4764	0.5388	0.5529
KUIDL-M-OPEN-1	0.4545	0.4647	0.5388	0.5529
TTOKU-D-ORCL-1	0.2915	0.3906	0.3123	0.4211
TTOKU-M-ORCL-1	0.1800	0.1957	0.2317	0.2531
TTOKU-M-ORCL-2	0.1111	0.1253	0.1945	0.2142
TTOKU-D-ORCL-2	0.0977	0.1278	0.2487	0.2883

"space-limited" and "time-limited" tasks like 1CLICK.

In contrast to Figure 17 which discusses per-run performances, Figure 18 provides a micro-level comparison of S-measure, readability and trustworthiness. First, per-query readability and trustworthiness values were averaged across the two assessors, which produced values ranging from -2 to 2. Next, for each average readability/trustworthiness pair, the corresponding queries were identified and the S-measure values were averaged across that set of queries. Although there are outliers (i.e., the three peaks in the figure), the general trend appears to be that (i) if both readability and trustworthiness values are low, then the S-measure is also low; and that (ii) if both readability and trustworthiness values are high, then the S-measure is also high.

Finally, a grain of salt: the Cohen's Kappa values between two assessors were quite low, both for readability and trustworthiness. The Kappa for readability varied between 0.03 and 0.65; that for trustworthiness varied between 0.02 and 0.62. These results suggest that these two criteria are neither well-defined nor well-understood. Thus the analysis reported in this subsection should be taken as preliminary.

5.6 Assessor Effort

Our current evaluation methodology relies on manual nugget match evaluation by means of a dedicated interface, and we do not attempt to automate this process at this stage, as we believe that simple string matching between nuggets and the system output will not be able to evaluate highly abstractive systems that present concise information. Thus it is very important that the nugget match evaluation can be completed in a reasonable time.

Figure 21 plots the average time spent for the nugget match evaluation for one query plotted against the number of nugget of that query. The average was taken over the 20 evaluations (10 X-strings times 2 assessors), except for Query 0033, for which the average was taken over 18 (9 X-strings times 2 assessors) because one run (TTOKU-D-ORCL-1) contained a formatting error (a missing TAB after the OUT field). The queries for which over 300 seconds were spent on average are indicated with a baloon: it can be observed that these are all CELEBRITY queries. These queries require large lists of nuggets to be returned. Thus, if we avoid such kind of queries, it would be possible to reduce the assessor effort substantially.

It can be observed that the average assessment time is highly correlated with the number of nuggets (Pearson correlation: .809). As was mentioned earlier, the total time spent by an assessor for assessing 120 X strings was 302 minutes on average. Thus the average time required for an assessor to evaluate and X-string was 151 seconds. We believe that our evaluation methodology is feasible, especially if queries are selected so that unnecessarily long lists of nuggets are avoided, as was mentioned above.

6. Conclusions and Future Work

Although the first round of the Once Click Access task had only three participating teams, they pursued diverse approaches: information extraction, passage retrieval and abstractive summarisation. The interested reader is encouraged to read the participant papers from KUIDL[6], MSRA1click[7] and TTOKU[8].

Our new evaluation framework, which involves manual nugget match evaluation and computation of S-measure, seems both feasible and useful for building systems that can present important information first within a small text window and minimise the amount of text the user has to read. Our experiments suggest that S-measure (a ranked retrieval metric for nuggets) and weighted recall (a set retrieval metric for nuggets) are comparable in terms of discriminative power but sometimes disagree as to which run pairs are significantly different. We view this as a good thing – that is the whole point of introducing a new metric that represents new task requirements.

We would like to extend 1CLICK to other languages such as English and Chinese at NTCIR-10. We should also probably reconsider the query types and the types of nuggets from the viewpoint of practical usefulness: for example, avoiding long lists of titles etc. for CELEBRITY queries may be both practical and cost-saving. Some more open problems are discussed in our CIKM paper [5].

7. Acknowledgments

We thank the NTCIR-9 1CLICK participants for their effort in producing the runs. In particular, we are indebted to the following people who participated in the nugget match evaluation: Takuya Makino, Hajime Morita, Hiroaki Ohshima, Naoki Orii, Yoshiyuki Shoji, Kosetsu Tsukuda, Takehiro Yamamoto, and Meng Zhao.

We also thank Miho Sugimoto and Noriko Kando for their support, especially in helping us devise the training topics. Finally, we thank the INTENT organisers and the NTCIR chairs for fruitful discussions.

8. References

[1] B. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*, to appear.

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

[3] T. Sakai. NTCIREVAL: A Generic Toolkit for Information Access Evaluation. In *Proceedings of the Forum on Information Technology 2011*, Vol. 2, pages 23–30, 2011.

[4] T. Sakai, D. Ishikawa, N. Kando, Y. Seki, K. Kuriyama and C.-Y. Lin. Using graded-relevance metrics for evaluating community QA answer selection. In *Proceedings of ACM WSDM 2011*, pages 187–196, 2011.

[5] T. Sakai, M. P. Kato and Y.-I. Song. Click the Search Button and Be Happy: Evaluating Direct and Immediate Information Access. In *Proceedings of ACM CIKM 2011*, pages 621–630, 2011.

[6] M. P. Kato, M. Zhao, K. Tsukuda, Y. Shoji, T. Yamamoto, H. Ohshima and K. Tanaka. Information Extraction based Approach for the NTCIR-9 1CLICK Task. In *Proceedings of NTCIR-9*, to appear, 2011.

[7] N. Orii, Y.-I. Song and T. Sakai. Microsoft Research Asia at the NTCIR-9 1CLICK Task. In *Proceedings of NTCIR-9*, to appear, 2011.

[8] H. Morita, T. Makino, T. Sakai, H. Takamura and M. Okumura. TTOKU Summarization Based Systems at NTCIR-9 1CLICK task. In *Proceedings of NTCIR-9*, to appear, 2011.

Table 6. Mean S-measure (I) vs. mean readability vs mean trustworthiness. The highest value in each column is shown in bold.

	S-measure (**I**)	readability	trustworthiness
KUIDL-D-OPEN-1	**0.3132**	0.37	**0.53**
KUIDL-D-OPEN-2	0.2900	0.34	0.46
KUIDL-M-OPEN-1	0.2196	0.58	0.35
KUIDL-M-OPEN-2	0.2214	**0.63**	0.37
MSRA1click-D-OPEN-1	0.2832	0.27	0.31
MSRA1click-D-OPEN-2	0.2988	0.28	0.42
TTOKU-D-ORCL-1	0.1585	−0.75	0.03
TTOKU-D-ORCL-2	0.1484	0.47	0.43
TTOKU-M-ORCL-1	0.0866	−0.83	−0.43
TTOKU-M-ORCL-2	0.0829	0.03	−0.07

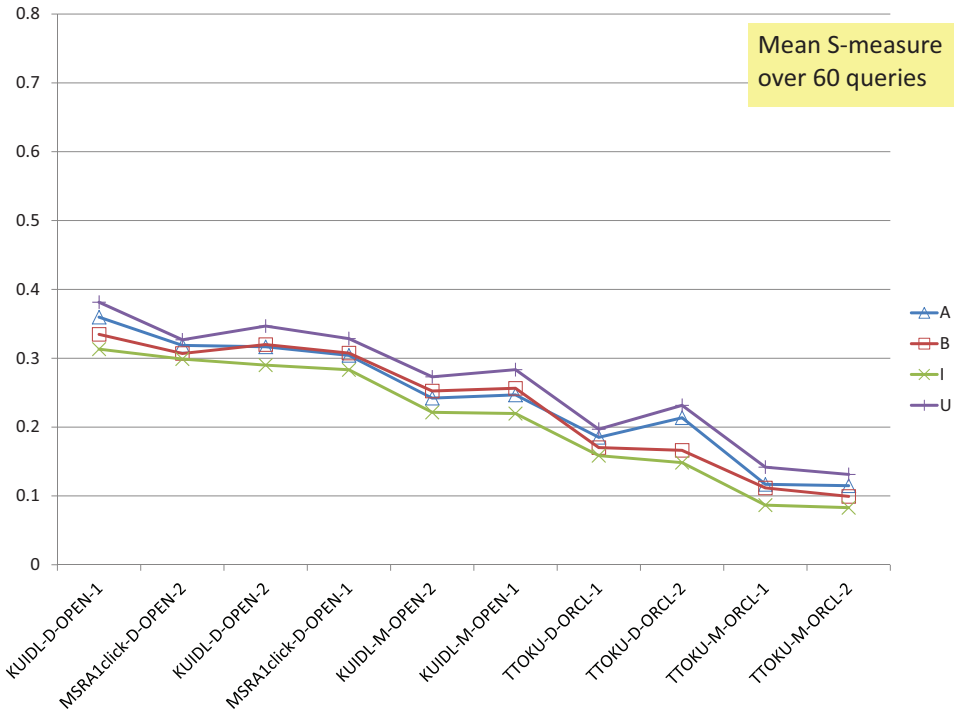

Figure 5. Mean S-measure: runs sorted by results based on I.

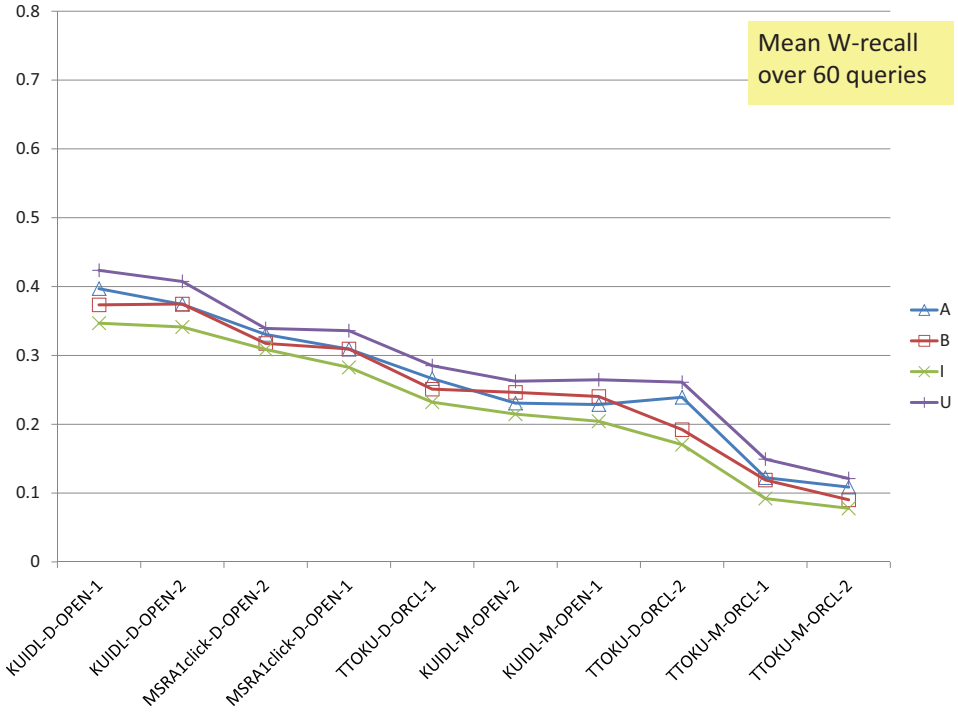

Figure 6. Mean W-recall: runs sorted by results based on I.

```
KUIDL-D-OPEN-1 with KUIDL-M-OPEN-1(W),KUIDL-M-OPEN-2(W),TTOKU-D-ORCL-1(S),TTOKU-D-ORCL-2(S,W),
                    TTOKU-M-ORCL-1(S,W),TTOKU-M-ORCL-2(S,W)
KUIDL-D-OPEN-2 with KUIDL-M-OPEN-1(W),KUIDL-M-OPEN-2(W),TTOKU-D-ORCL-1(S),TTOKU-D-ORCL-2(S,W),
                    TTOKU-M-ORCL-1(S,W),TTOKU-M-ORCL-2(S,W)
KUIDL-M-OPEN-1 with TTOKU-M-ORCL-1(S),TTOKU-M-ORCL-2(S,W)
KUIDL-M-OPEN-2 with TTOKU-M-ORCL-1(S),TTOKU-M-ORCL-2(S,W)
MSRA1click-D-OPEN-1 with TTOKU-D-ORCL-1(S),TTOKU-D-ORCL-2(S),TTOKU-M-ORCL-1(S,W),TTOKU-M-ORCL-2(S,W)
MSRA1click-D-OPEN-2 with TTOKU-D-ORCL-1(S),TTOKU-D-ORCL-2(S,W),TTOKU-M-ORCL-1(S,W),TTOKU-M-ORCL-2(S,W)
TTOKU-D-ORCL-1 with TTOKU-M-ORCL-1(W),TTOKU-M-ORCL-2(W)
```

Figure 7. Significantly different run pairs in terms of S-measure and/or W-recall with I (randomised Tukey's HSD at $\alpha = 0.05$). "(S)" means only S-measure detected a significant difference; "(W)" means only W-recall detected a significant difference; "(S,W)" means both metrics detected a significant difference.

```
KUIDL-D-OPEN-1 with KUIDL-M-OPEN-1(W),KUIDL-M-OPEN-2(W),TTOKU-D-ORCL-1(S,W),TTOKU-D-ORCL-2(S,W),
                    TTOKU-M-ORCL-1(S,W),TTOKU-M-ORCL-2(S,W)
KUIDL-D-OPEN-2 with KUIDL-M-OPEN-1(W),KUIDL-M-OPEN-2(W),TTOKU-D-ORCL-1(S),TTOKU-D-ORCL-2(S,W),
                    TTOKU-M-ORCL-1(S,W),TTOKU-M-ORCL-2(S,W)
KUIDL-M-OPEN-1 with TTOKU-M-ORCL-1(S),TTOKU-M-ORCL-2(S,W)
KUIDL-M-OPEN-2 with TTOKU-M-ORCL-1(S),TTOKU-M-ORCL-2(S,W)
MSRA1click-D-OPEN-1 with TTOKU-D-ORCL-1(S),TTOKU-M-ORCL-1(S,W),TTOKU-M-ORCL-2(S,W)
MSRA1click-D-OPEN-2 with TTOKU-D-ORCL-1(S),TTOKU-M-ORCL-1(S,W),TTOKU-M-ORCL-2(S,W)
TTOKU-D-ORCL-1 with TTOKU-M-ORCL-1(W),TTOKU-M-ORCL-2(W)
TTOKU-D-ORCL-2 with TTOKU-M-ORCL-2(W)
```

Figure 8. Significantly different run pairs in terms of S-measure and/or W-recall with U (randomised Tukey's HSD at $\alpha = 0.05$). "(S)" means only S-measure detected a significant difference; "(W)" means only W-recall detected a significant difference; "(S,W)" means both metrics detected a significant difference.

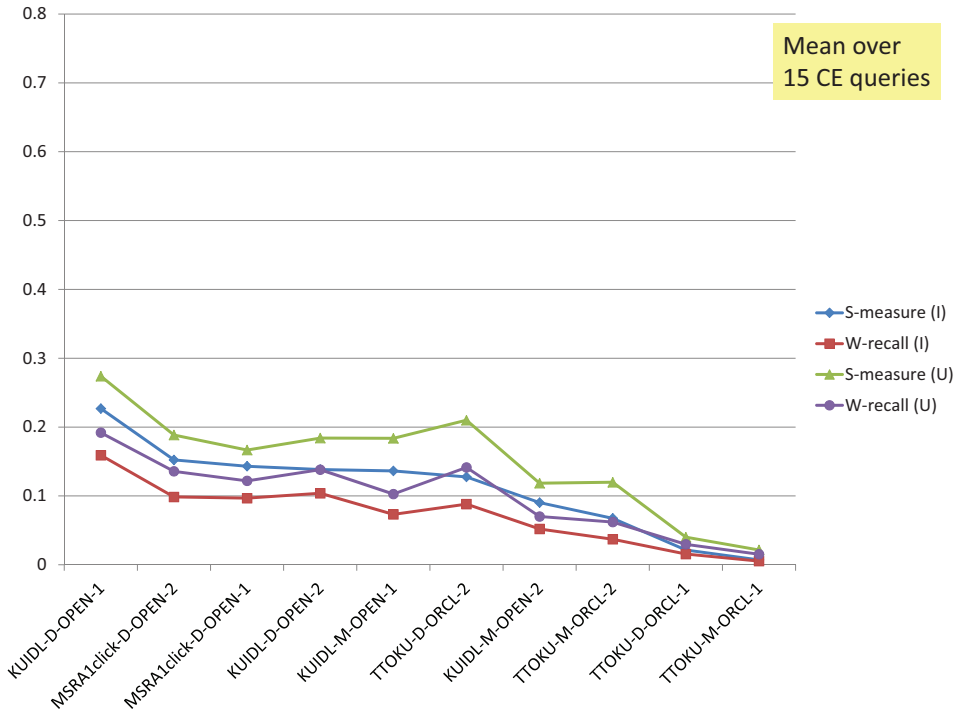

Figure 9. Mean over CE queries: runs sorted by Mean S-measure with I.

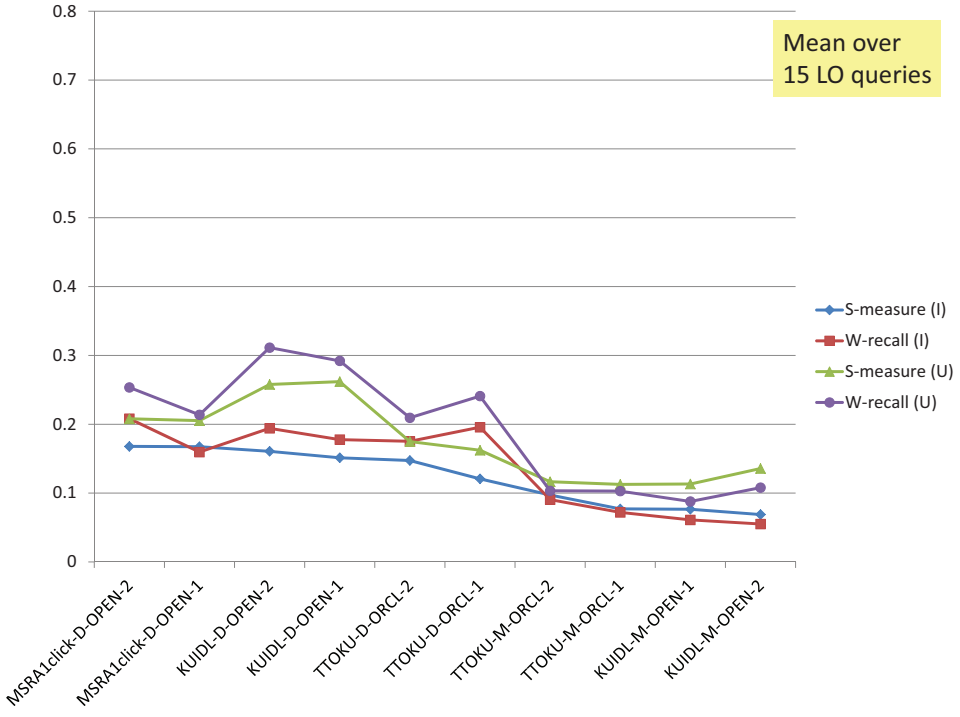

Figure 10. Mean over LO queries: runs sorted by Mean S-measure with I.

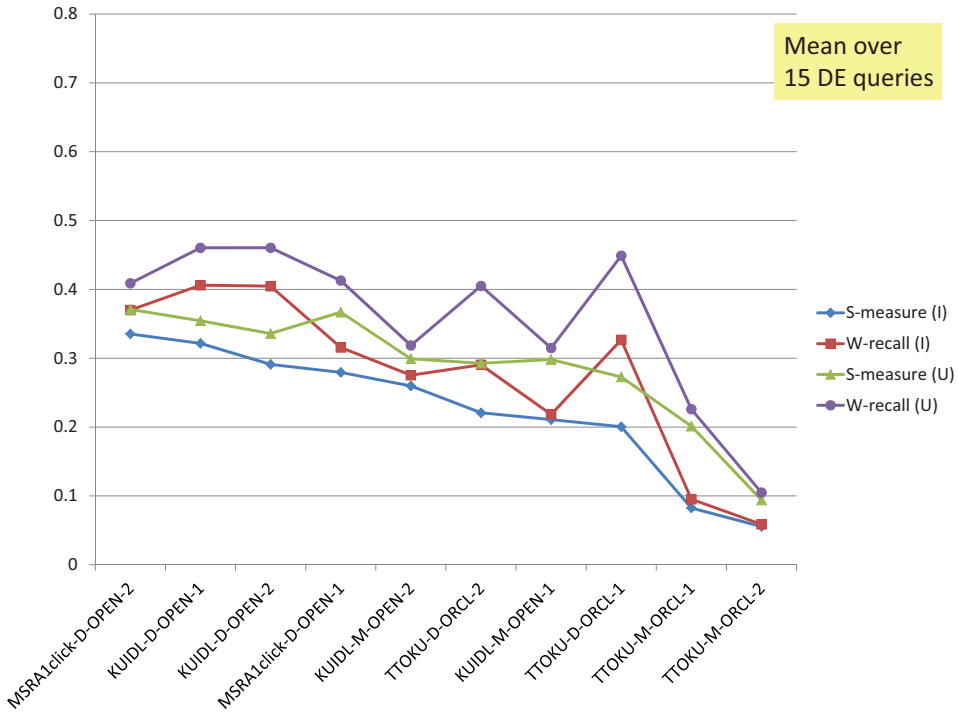

Figure 11. Mean over DE queries: runs sorted by Mean S-measure with I.

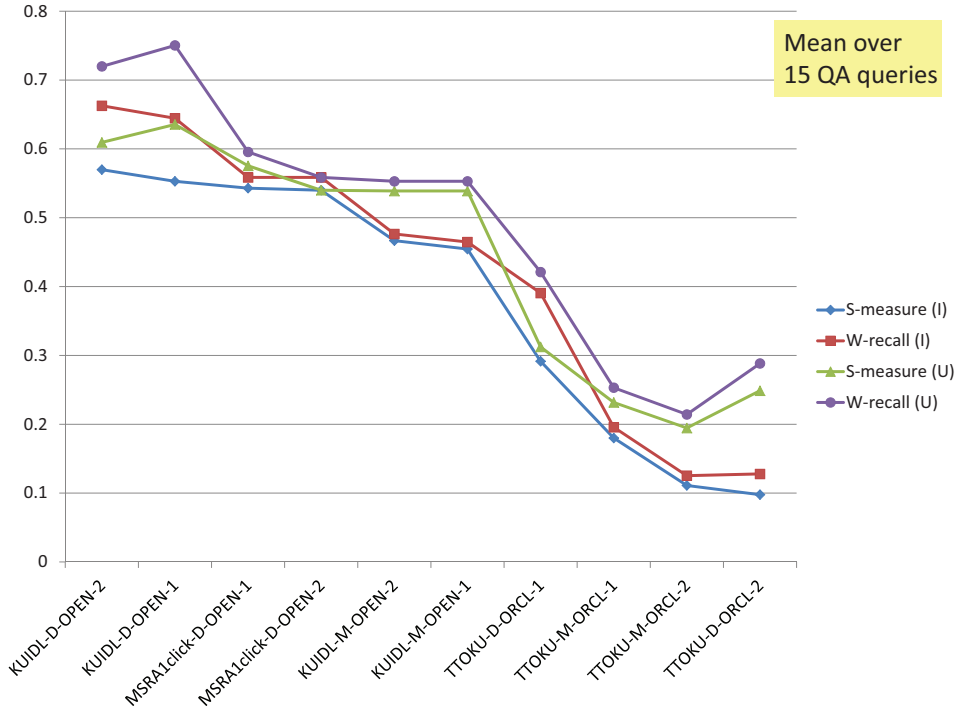

Figure 12. Mean over QA queries: runs sorted by Mean S-measure with I.

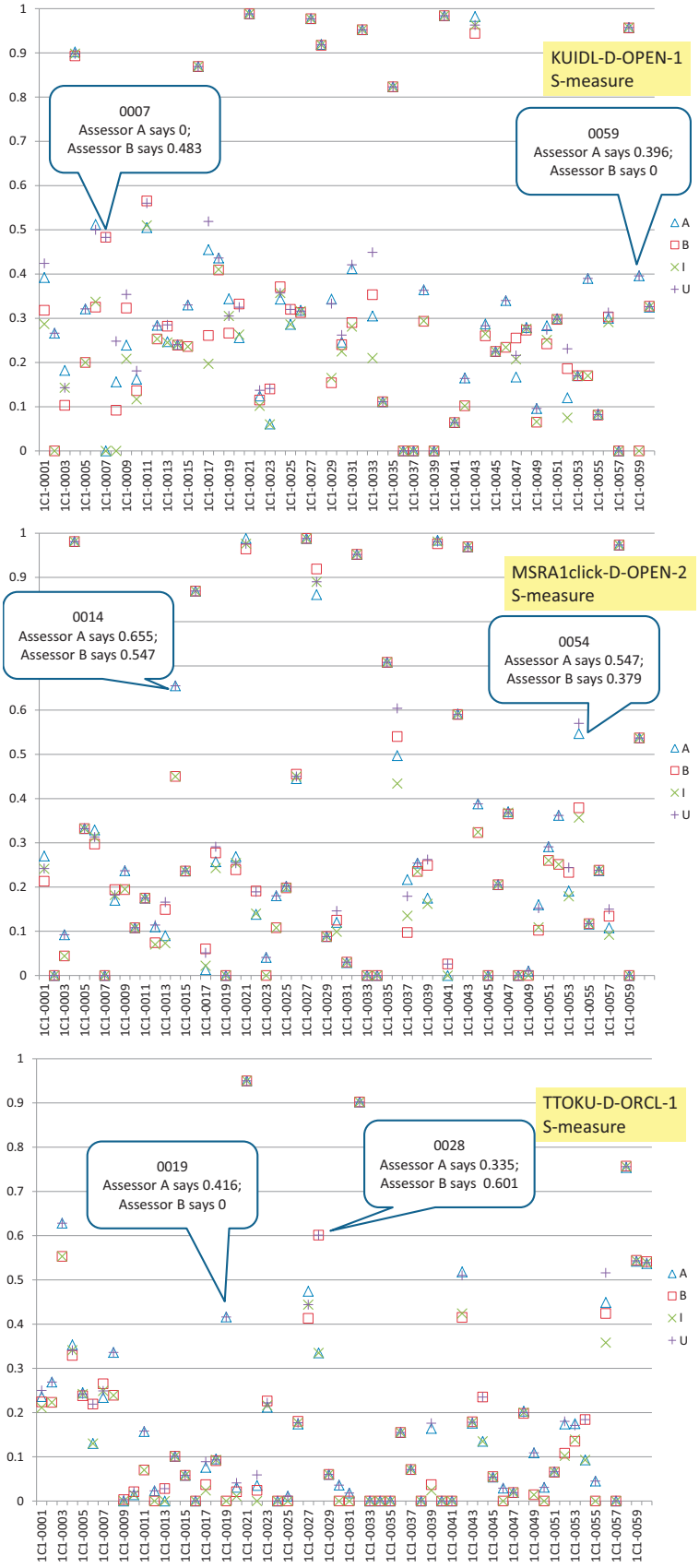

Figure 13. Per-query inter-assessor disagreement in terms of S-measure.

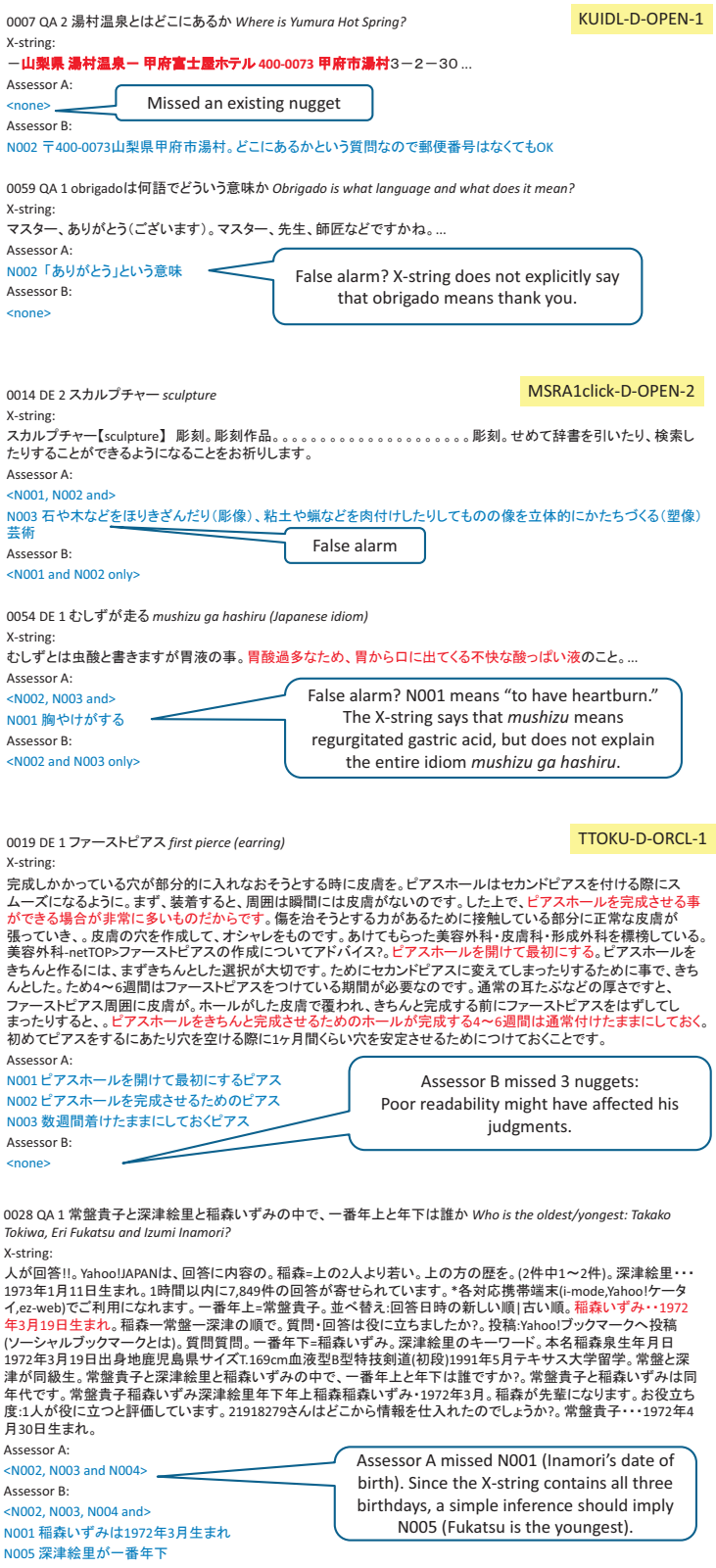

Figure 14. Investigation of inter-assessor differences.

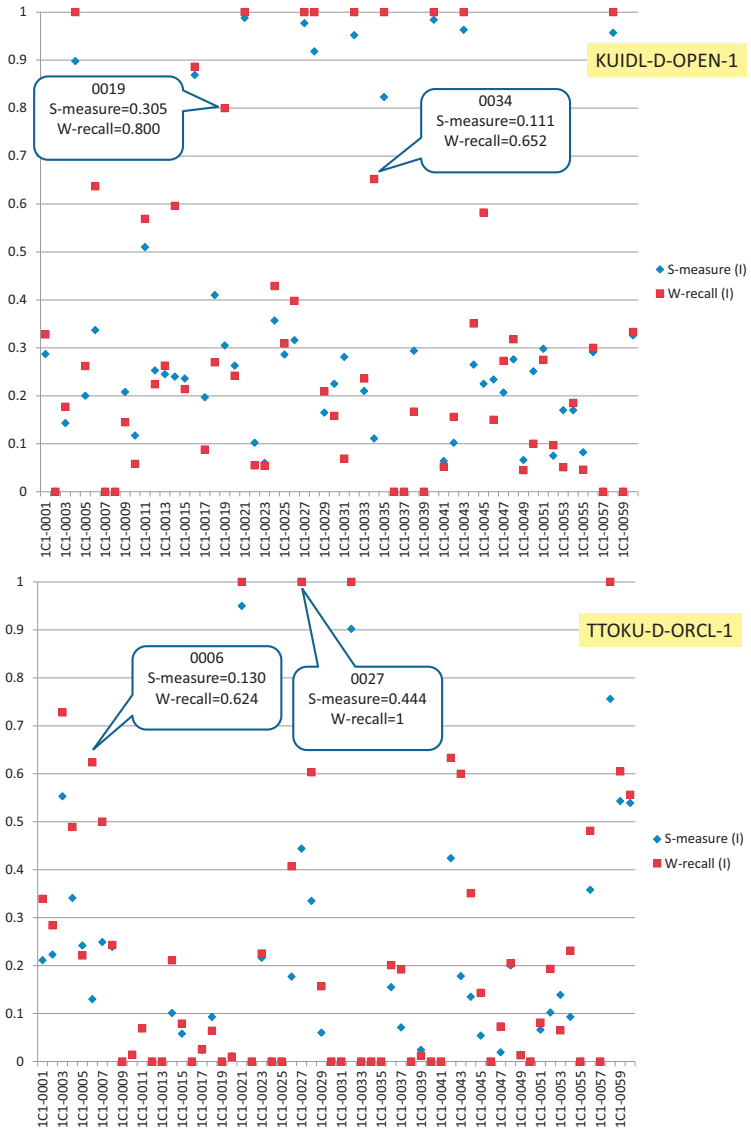

Figure 15. Per-query disagreement between S-measure and W-recall.

0019 DE 1 ファーストピアス *first pierce (earring)* KUIDL-D-OPEN-1

X-string:

今から始めるピアスデビュー。スポンサードリンク。ストレートバーベルの18Gがいいと思いますよ。thecooltrader。Pagetop。Copyrightcバスローブバスローブを使おう。Allrightsreserved。エンジニアブーツ｜。｜ビタクラフトの口コミ｜目覚まし時計｜杖ステッキ。insertedbyFC2system。モンスターハンター3rdWikiマニアック最新ニュース;にんにく玉本舗のにんにく卵黄、肝臓強化で二日酔いににんにく玉本舗のにんにく卵黄、肝臓強化で二日酔いに;joueravecmoa。穴を開けてから1ヶ月ほど付けておくことで、ピアス穴を完成させます。ファーストピアスを知ろう。初めてピアス穴を作るために使用するピアスのことを言います。最初に付けるピアスのこと。この穴が十分出ないと、すぐに閉じてしまったり、ピアストラブルを起こすことになります。ファーストピアスの材質は24金かチタンが適している。ファーストピアスって。]ファーストピアスについて解説しています。ピアスホールが、完成するまで4～6週間は、。2007年8月2日。

N001 ピアスホールを開けて最初にするピアス
N002 ピアスホールを完成させるためのピアス
N003 数週間着けたままにしておくピアス

3 out of 4 nuggets are found, but the first half of the X-string is mostly irrelevant (mentions barbells, bathrobes, engineer boots, alarm clocks, garlic etc.)

0034 QA 1 ANA国内線で機内に持ち込める手荷物のサイズ *Size of carry-on luggage allowed on ANA domestic flights*

X-string:

満席の場合、全員が制限いっぱいの荷物を持っていたら収納しきれない可能性があります。また、機種にもより座席により、頭上収納が使えないところもあります。その場合はCAさんが預かってキャビネットなどにしまってくれます。安心してください。...そのスーツケースは機内に持ち込むのでしょうか？搭乗手続き時に預けるのであれば何ら問題ありません。機内に持ち込むのであれば、ハンドバッグが小さめのものであればOKです。（機内に持ち込む荷物は1人1個までですが、身の回りの品は1個と数えない特例があります...添付された画像のサイズ表記が、ハンドルや車輪も含めた正しいサイズである、という前提で回答させて頂きます。サイズが49×35×17cmですので、機内に持ち込めるかどうかは、ご利用になる航空機の座席数で分かれます。座席数100席未満の航空機(ANAの時刻...飛行機の座席数によって異なるようです。(国内線共通）座席数が100席以上 3辺の合計が115cm以内(55cm×40cm×25cm以内) 100席未満 3辺の合計が100cm以内(45cm×35cm×20cm以内) http://www.ana.c.

N001 100席以上の機種では3辺の合計が115cm以内(55cm×40cm×25cm以内)
N003 100席未満の機種では3辺の合計が100cm以内(45cm×35cm×20cm以内)

Relevant part is at the end of the X-string

0006 LO 1 神戸市立中央図書館 TTOKU-D-ORCL-1

詳しい地図はこちらへ。中央図書館への道順・地図。館内整理日(年4回、2・6・10・12月の第3木曜)。サイトマップ神戸市ホームページのご利用案内プライバシーポリシー。トップページ利用案内施設案内・アクセス開館時間と休館日行事調べ物ガイド新聞・雑誌所蔵一覧神戸ふるさと文庫1.17文庫(震災関連資料)KOBEの本棚事業概要書燈(神戸市立図書館報)意見募集(パブリックコメント)図書館協議会アンケート東灘図書館移転整備子どものページおうちのかた・先生がたへえ小箱。ページの内容についてメールで問い合わせする。サイトマップお問い合わせよくある質問と回答。神戸市トップページへページの先頭へ。ページの作成責任者は、総務課長小林史明です。市政やくらし、イベント情報などのお問い合わせは、神戸市総合コールセンターまで。〒650-0017神戸市中央区楠町7丁目2-1中央体育館・文化ホール山側/大倉山公園内Tel:078-371-3351Fax:078-371-5046交通地下鉄大倉山駅北200m高速神戸駅北500mJR神戸駅北800m。毎週月曜(祝日・休日の場合は開館、直後の祝日・休日でない日を休館)。

N003 467 休館日 毎週月曜（祝日・休日の場合は開館、直後の祝日・休日でない日を休館）。毎週月曜だけではNG
N007 353 住所〒650-0017神戸市中央区楠町7丁目2-1
N008 388 phone 078-371-3351
N009 403 fax 078-371-5046
N010 417 アクセス 地下鉄大倉山駅 北200m
N011 427 アクセス 高速神戸駅 北500m
N012 437 アクセス JR神戸駅 北800m

Relevant part is at the end of the X-string

0027 DE 1 日本国民の三大義務

人権の歴史的性格と、保持のために必要な国民の責務をうたったもので、。国民は、これを濫用してはならないのであつて、。国民主権国家においては、国民の納める税金によつてのみ国家の財政が維持され、。国民にとつての精神的指針という意味が大きく、法的義務を課した規定ではない。常に公共の福祉のためにこれを利用する責任を負ふ。国家の存立と国政の運営が可能となることからして、国民の当然の義務と解される。国民の義務について述べよ。2教育の義務について形式的には国家に対するものであるが、実質的には保護する子女に対するものである。第二六条すべて国民は、法律の定めるところにより、。A自由・権利保持の義務。「概念」一般的義務規定(12)、教育の義務(26)、勤労の義務(27)、納税の義務(30)。3勤労の義務について。保護する子女に普通教育を受けさせる義務を負ふ。B自由・権利を濫用しない義務。第一二条憲法が国民に保障する自由及び権利は、。第二七条すべて国民は、勤労の権利を有し、義務を負ふ。C自由・権利を公共の福祉のために利用する義務、に分けられる。国民の不断の努力によつて、これを保持しなければならない。4納税の義務について。第三〇条国民は、法律の定めるところにより、納税の義務を負ふ。

N001 243 教育の義務
N002 295 勤労の義務
N003 302 納税の義務

3 out of 3 nuggets are found, but the first half of the X-string is not relevant (though not off-topic).

Figure 16. Investigation of inter-metric differences.

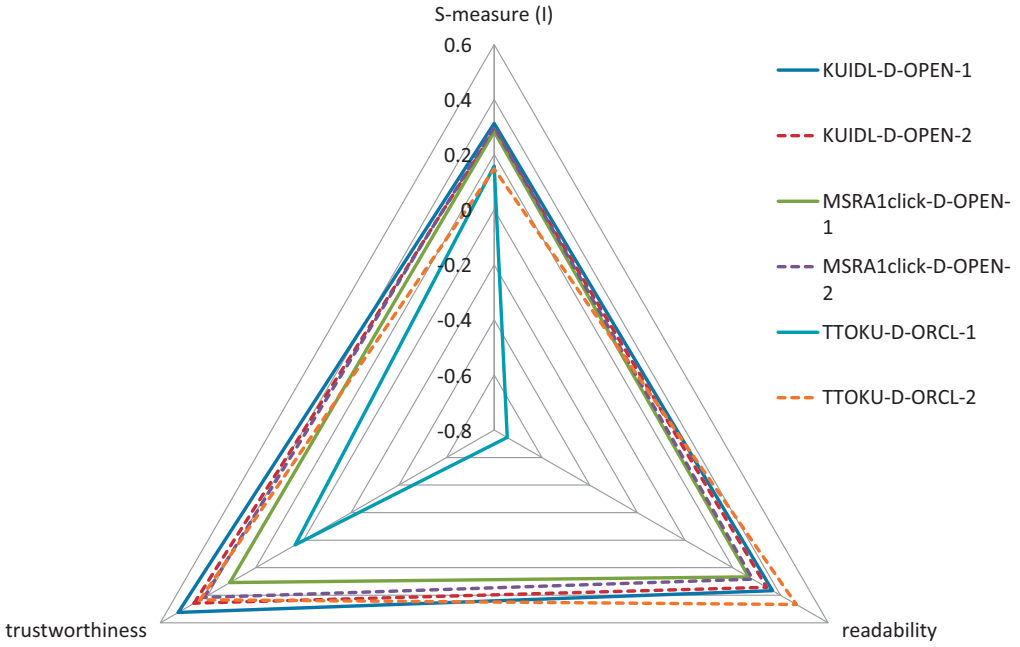

Figure 17. Mean S-measure (I) vs. mean readability vs. mean trustworthiness for the D-runs.

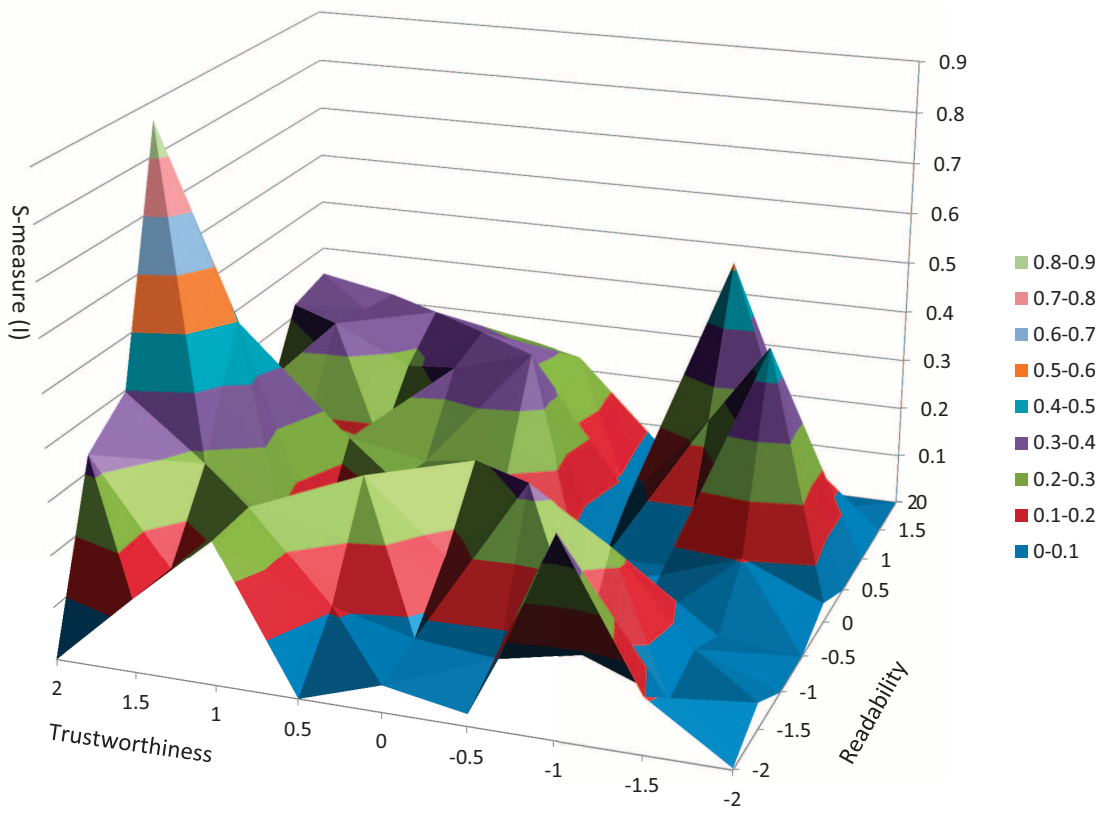

Figure 18. Average S-measure (I) vs. average readability vs. average trustworthiness. The average is taken over the two assessors.

KUIDL-D-OPEN-1	1C1-0006	FAXに誤りが見られる．番号からしてまったく別の住所であると思われる．
KUIDL-D-OPEN-1	1C1-0017	文字化け
KUIDL-D-OPEN-1	1C1-0017	ナゲット４・５に対応していそうな箇所が読めず．営業時間:月曜縲恚燉j 午前8時45分縲怩゛ 後5時15分 第2・4土曜 午前9時00分縲協ウ午(戸籍課、保険年金課、こども家庭支援課(一部業務)の窓口)。
KUIDL-D-OPEN-1	1C1-0029	「大江戸線春日駅徒歩5分」の「徒歩3分」は誤差とした．ナゲットに時間が含まれるが，本文中に含まれないケースが多い．
KUIDL-D-OPEN-1	1C1-0033	両方の情報が入り混じっているから実際には使い物にならない。
KUIDL-D-OPEN-1	1C1-0033	13:車では入っていないが，インターチェンジから徒歩などは考えづらいので正解とした．
KUIDL-D-OPEN-1	1C1-0042	この文章もほぼ同じ内容が繰り返されている．情報量は少ないが，読みやすくはある．また，前半がBlogのような内容であるため，Trustworthinessはやや低い．
KUIDL-D-OPEN-1	1C1-0044	ほとんど同じ意味のフレーズが繰り返し用いられている．文書的には読みやすい．
KUIDL-D-OPEN-1	1C1-0046	整体あんの情報が混ざっているため，Trustworthinessは低いと判断した．
KUIDL-D-OPEN-2	1C1-0001	「利用可能時間:9:00-16:30(季節変動あり)」を「開園時間9～16:50分」とマッチさせるかどうか．適合と判断した．
KUIDL-D-OPEN-2	1C1-0003	別のLocationに関する情報が多々ある．
KUIDL-D-OPEN-2	1C1-0017	ナゲット４・５に当たる箇所が文字化けで読めず．
KUIDL-D-OPEN-2	1C1-0020	52が文字化け
KUIDL-D-OPEN-2	1C1-0033	2つの異なる福岡工業高等学校の情報が混ざっていて読みづらい．
KUIDL-D-OPEN-2	1C1-0033	福岡と岩手の情報が混在
KUIDL-D-OPEN-2	1C1-0036	ホテルサンライン蒲田のじょうほう
KUIDL-D-OPEN-2	1C1-0052	乗っている営業時間や住所、電話番号は大丸内のテナントののやのもの。
KUIDL-D-OPEN-2	1C1-0058	2行目の「グーグルはヤフーやグーなどのほかの検索エンジン」は「グーグルは検索サイト（検索エンジン）である」の正解として良いか．
KUIDL-M-OPEN-1	1C1-0023	ナゲットの不完全なカバーが多い。たとえば、ロシア人であることが読み取れない
KUIDL-M-OPEN-1	1C1-0030	政治家であることは「所属政党」という記述から．「衆議院議員当選回数 8回」に関しては，衆議院という語がないため，マッチしないとした．
KUIDL-M-OPEN-1	1C1-0030	祖父、叔父、中川義雄、中川文蔵、中川正男　という表記では、文蔵がどっちなのかよくわからない
KUIDL-M-OPEN-1	1C1-0031	「代表作」はアニメでなく漫画と解釈した。
KUIDL-M-OPEN-1	1C1-0033	福岡の情報を岩手の情報が実は混在しておりmisleading
KUIDL-M-OPEN-1	1C1-0036	誤情報が多数見られる．
KUIDL-M-OPEN-2	1C1-0003	誤った情報を多数含んでいる．
KUIDL-M-OPEN-2	1C1-0006	別の情報が混ざっている．
KUIDL-M-OPEN-2	1C1-0007	2のnuggetは，回答文では甲府市湯村は甲府富士屋ホテルの場所と読めるため
KUIDL-M-OPEN-2	1C1-0008	誤情報を含む
KUIDL-M-OPEN-2	1C1-0032	だそうです　がこころもとない
KUIDL-M-OPEN-2	1C1-0033	11.の回答中に文字化け有り
KUIDL-M-OPEN-2	1C1-0036	ホテルサンライン蒲田の情報です。
KUIDL-M-OPEN-2	1C1-0043	正解に　勘解由小路　が入っていない
KUIDL-M-OPEN-2	1C1-0052	本文中の「営業時間:10時~20時」はあるテナント（ののや）のもの
KUIDL-M-OPEN-2	1C1-0057	Gacktとは何だよ

Figure 19. Comments obtained through nugget match evaluation: part I.

MSRA1click-D-OPEN-1	1C1-0007	六甲山や神戸市内という地名から推測することもできるが位置は特定できないため不適合とした.
MSRA1click-D-OPEN-1	1C1-0010	生年月日2005年(h17) 1月20日がおかしい.
MSRA1click-D-OPEN-1	1C1-0012	なかにし礼？
MSRA1click-D-OPEN-1	1C1-0018	4つ目のnuggetは「所属している芸能事務所」からタレントとわかると判断.
MSRA1click-D-OPEN-1	1C1-0025	矛盾：生年月日1981年4月14日生年月日(2008年10月05日)
MSRA1click-D-OPEN-1	1C1-0026	矛盾：ドーハはイランの都市ではありません。ドーハはイランの都市でもあります。
MSRA1click-D-OPEN-1	1C1-0029	地下鉄後楽園駅から徒歩2分あしがおそい
MSRA1click-D-OPEN-1	1C1-0030	正解の選挙区　(北海道第5区→) 北海道第11区は両方？
MSRA1click-D-OPEN-1	1C1-0043	勘解由小路　が正解にない
MSRA1click-D-OPEN-1	1C1-0048	肝心なところで途切れているため適合なしとした.
MSRA1click-D-OPEN-1	1C1-0049	まったく関係のない内容だと思われる.
MSRA1click-D-OPEN-1	1C1-0053	別人の情報が含まれ，生年月日が混同される可能性があるため，Trustworthinessは低いと判定した.
MSRA1click-D-OPEN-1	1C1-0053	京都市南区 が出身だとは書いてない
MSRA1click-D-OPEN-1	1C1-0058	Blog的な文章であるため，やはりTrustworthinessを低くした.
MSRA1click-D-OPEN-2	1C1-0003	トピックが入り混じり，文の境界がわからないため読みづらい.
MSRA1click-D-OPEN-2	1C1-0013	本田　朋子 (ほんだ　ともこ) 1983年8月16日生　愛媛県出身 だれ？
MSRA1click-D-OPEN-2	1C1-0016	宇都宮タワーの方が高いと思われるような表記が見られるため，低いTrustworthinessとした.
MSRA1click-D-OPEN-2	1C1-0026	矛盾:ドーハはイランの都市ではありません。ドーハはイランの都市でもあります。
MSRA1click-D-OPEN-2	1C1-0059	フットサル？
TTOKU-D-ORCL-1	1C1-0002	文章として成立しているとは考えにくい.
TTOKU-D-ORCL-1	1C1-0023	＞ミハイル・カラシニコフの最終階級は技術中将存命なので今後上がらないとも限らない
TTOKU-D-ORCL-1	1C1-0031	文章として成立していない。作品名が出現していても、それが漫画なのかアニメなのか、そもそも手塚治虫の作品なのかがこの文章からは分からない。
TTOKU-D-ORCL-1	1C1-0048	多くの文章で意味不明。
TTOKU-D-ORCL-1	1C1-0053	文章がところどころ組み合わされていて難解である.
TTOKU-D-ORCL-1	1C1-0054	日本語として文章がおかしいものが多い.
TTOKU-D-ORCL-1	1C1-0055	ワードサラダ。
TTOKU-D-ORCL-2	1C1-0010	いくつかの受賞に関して年度がないために不適合のものがあった.
TTOKU-D-ORCL-2	1C1-0014	スカルプチャがつけ爪だということは，1つのセンテンスを読んで初めて理解できる.
TTOKU-D-ORCL-2	1C1-0015	出力はSCCとNSEの説明ではないようである．ただし，内容はしっかりとしている.
TTOKU-D-ORCL-2	1C1-0030	明確には政治家と書かれていないが，選挙区，内閣，政党から類推可能．ナゲットがやや大きく，ナゲットをカバーできていない場合が多数見受けられた．e.g. ナゲットが年＋出来事の場合
TTOKU-D-ORCL-2	1C1-0031	文章として成立していない。作品名が出現していても、それが漫画なのかアニメなのか、そもそも手塚治虫の作品なのかがこの文章からは分からない。
TTOKU-D-ORCL-2	1C1-0048	空出力.
TTOKU-D-ORCL-2	1C1-0050	後半の羅列はそれが何なのか推測できない。
TTOKU-D-ORCL-2	1C1-0057	ジャンプや回転はお手の物???
TTOKU-D-ORCL-2	1C1-0059	obrigado, not origado!
TTOKU-M-ORCL-1	1C1-0004	やや判断が分かれるかもしれないが，熊猫が来園という事実から，王子動物園が答えであることを推測できる.
TTOKU-M-ORCL-1	1C1-0030	解読不能.
TTOKU-M-ORCL-1	1C1-0038	読めない.
TTOKU-M-ORCL-1	1C1-0051	盗作疑惑があるという意見としては貴重？
TTOKU-M-ORCL-2	1C1-0010	受賞したことが読み取れない
TTOKU-M-ORCL-2	1C1-0022	「真鍋を学べ！」が出版であることはISBNから推測できる．非常に断片的で読みづらい.
TTOKU-M-ORCL-2	1C1-0026	情報はあるが，単語のみで説明がないため，不適合としたものがいくつかある.
TTOKU-M-ORCL-2	1C1-0044	単語の羅列であるため，解読が難しい.

Figure 20. Comments obtained through nugget match evaluation: part II.

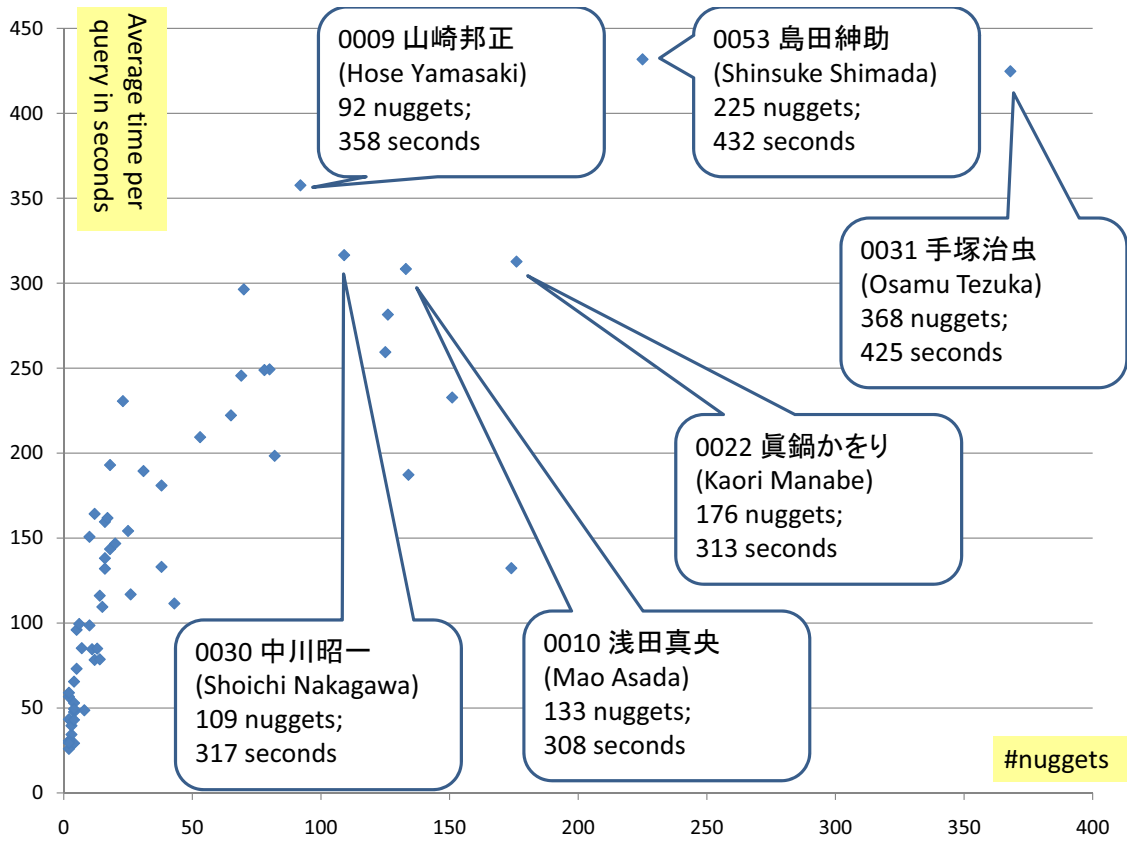

Figure 21. Nugget evaluation time per query plotted against the number of nuggets. "Time A" and "Time B" represents the time spent by "Assessors" A and B; the average over two assessors is also shown.