One Click One Revisited: Enhancing Evaluation based on Information Units

Tetsuya Sakai[1,a)] Makoto P. Kato[2,b)]

Abstract: This paper extends the evaluation framework of the NTCIR-9 One Click Access Task (1CLICK-1), which required systems to return a single, concise textual output in response to a query in order to satisfy the user immediately after a click on the SEARCH button. Unlike traditional nugget-based summarisation and question answering evaluation methods, S-measure, the official evaluation measure of 1CLICK-1, discounts the value of each information unit based on its position within the textual output. We first show that the discount parameter L of S-measure affects system ranking and discriminative power, and that using multiple values, e.g. $L = 250$ (user has only 30 seconds to view the text) and $L = 500$ (user has one minute), is beneficial. We then complement the recall-like S-measure with a simple, precision-like measure called T-measure as well as a combination of S-measure and T-measure, called $S\sharp$. We show that $S\sharp$ with a heavy emphasis on S-measure imposes an appropriate length penalty to 1CLICK-1 system outputs and yet achieves discriminative power that is comparable to S-measure. These new measures will be used at NTCIR-10 1CLICK-2.

Keywords: information units, NTCIR, One Click Access, S-measure, T-measure, $S\sharp$.

1. Introduction

The NTCIR-9 One Click Access Task ("1CLICK-1," pronounced *One Click One*) was concluded in December 2011 [18]. In contrast to traditional information retrieval (IR) and web search where systems output a ranked list of items in response to a query, 1CLICK-1 required systems to output one piece of concise text, typically a multi-document summary of several relevant web pages, that fits (say) a mobile phone screen. Participating systems were expected to output important pieces of information first, and to minimise the amount of text the user has to read in order to obtain the desired information. The task was named One Click Access because systems were required to satisfy the user immediately after the user issues a simple query and clicks on the SEARCH button. This task setting fits particularly well to a mobile scenario in which the user has very little time to interact with the system [17].

To go beyond *document* retrieval and design advanced *information* retrieval systems such as 1CLICK systems, the IR community needs to explore evaluation based on *information units* ("iUnits") rather than *document* relevance [1][*1]. An iUnit should be an atomic piece of information that stands alone and is useful to the user. At 1CLICK-1, *S-measure* was used to evaluate participating systems based on iUnits: this is a generalisation of the *weighted recall* of iUnits ("W-recall"), but unlike W-recall it takes the *positions* of retrieved iUnits into account.

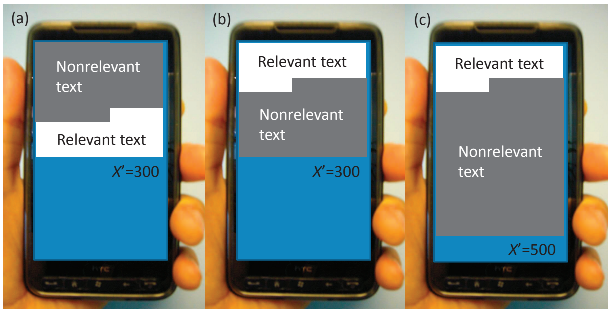


Fig. 1 X-strings: the output of 1CLICK systems.

Figure 1 shows a few conceptual images of texts output by 1CLICK systems, called *X-strings* as 1CLICK-1 systems were required to return a text whose target length is no more than X characters. The X-strings in Fig. 1(a) and (b) are both 300-character long ($X' = 300$, where X' is the *actual* length), and they contain exactly the same pieces information that are relevant to a particular query. However, while the X-string in (a) makes the user read some nonrelevant text before he can get to the relevant text, that in (b) shows the same relevant text first. In this sense, the user can reach the desired information more efficiently with (b) than with (a). While W-recall and traditional "nugget-based" evaluation measures in summarisation and question answering regard (a) and (b) as equally effective, S-measure rewards (b) more heavily than (a). This *position-sensitive* evaluation can help researchers design effective 1CLICK systems.

S-measure has a parameter called L which represents the user's patience: at 1CLICK-1, where a *Japanese* task was evaluated, L

[1] Microsoft Research Asia, China
[2] Kyoto University, Japan
[a)] tetsuyasakai@acm.org
[b)] kato@dl.kuis.kyoto-u.ac.jp

[*1] We distinguish the iUnits in the 1CLICK evaluation framework from nuggets used in summarisation and question answering evaluation. As discussed later, the key difference between an iUnit and a traditional nugget is that the former contains *vital string* information [17], [18], as we shall explain later.

was set to 500 based on the statistic that the average reading speed of a Japanese is 400-600 characters per minute. Thus $L = 500$ implies that the user has only *one minute* to gather the desired pieces of information. The first objective of the present study is to examine the effect of L on the evaluation outcome of the participating systems at 1CLICK-1. For example, suppose the user only has *thirty seconds* to read the X-string: would the official system rankings change?

The second objective is to complement the official evaluation reported at 1CLICK-1, by proposing a simple extension to the iUnit-based evaluation. Compare Fig. 1(b) and (c): the two X-strings contain the same relevant information in the same positions, but the one in (c) contains more nonrelevant text: it makes the user waste more time. However, as S-measure is a position-sensitive version of W-recall, it cannot differentiate between (b) and (c). Hence we introduce a precision-like measure called *T-measure* and a combination of S-measure and T-measure (or "S" and "T" for short) called $S\sharp$, and demonstrate that they provide new insight into the 1CLICK-1 systems.

2. Related Work

2.1 Evaluating Search

The present study builds on the S-measure framework for evaluating 1CLICK-1 systems [17], [18]. There is an analogy between the well-known *normalised Discounted Cumulative Gain* (nDCG) [7] and S: while nDCG evaluates a *ranked list of items* (e.g. URLs) while discounting the value of each item based on their *rank positions*, S evaluates a *textual output* (i.e. the X-string) while discounting the value of each iUnit based on their *offset positions* within the output, to reward systems that satisfy the user quickly. S assumes that the user's reading speed is constant, and therefore that the discounting function is linear, unlike the case with nDCG.

For evaluating IR from the viewpoint of time spent to reach the desired information, Dunlop proposed *expected search duration* fifteen years ago [6]. A more popular and recent measure is *α-nDCG* designed primarily for diversified IR evaluation, which views both documents and search intents as sets of nuggets [5]. But these measures are for a ranked list of documents. The INEX Snippet Retrieval track[*2] evaluates the quality of *snippets* as a means to judge the relevance of the original documents within the traditional ranked list evaluation framework.

Recently, Pavlu *et al.* [13] have proposed a nugget-based evaluation framework for IR that involves automatic matching between documents and gold-standard nuggets. They are now jointly running the NTCIR-10 1CLICK-2 task with Sakai, Kato and Song [17], [18] to explore evaluation approaches based on iUnits[*3].

2.2 Evaluating Summarisation

ROUGE is a family of measures for evaluating summaries *automatically* [9]. The key idea is to compare a system output with a set of gold-standard summaries in terms of *recall* (or alternatively *F-measure*), where recall is defined based on automatically extracted textual fragments such as N-grams and longest common subsequences. New automatic summarisation measures were also explored at the TAC (Text Analysis Conference) AESOP (Automatically Evaluating Summaries of Peers) task[*4].

While automatic evaluation methods such as ROUGE are useful for efficient evaluation of summarisers, the S-measure framework builds on the view that automatic string matching between the system output and gold standards is not sufficient for building effective *abstractive* summarisers [17]. Thus, in the S-measure framework, the identification of iUnits within an X-string is done manually. More importantly, the assessor records the *position* of each iUnit. As we discussed earlier, this enables the S-measure framework to distinguish between systems like Fig. 1(a) and (b).

The S-measure framework is similar to the *pyramid method* for summarisation evaluation [12] in that it relies on manual matching. In the pyramid method, *Semantic Content Units* (SCUs) are extracted from multiple gold-standard summaries, and each SCU is weighted according to the number of gold standards it matches with. Finally, SCU-based weighted precision or recall is computed. Just like the automatic methods, however, these methods are *position insensitive*.

2.3 Evaluating Distillation

The DARPA GALE *distillation* program evaluated ranked lists of passages output in response to a query (or a set of queries representing a long-standing information need). Within this framework, Babko-Malaya [3] describes a systematic way to define nuggets in a bottom-up manner from a pool of system output texts. In contrast, the iUnits were defined prior to run submissions at 1CLICK-1 [18].

White, Hunter and Goldstein [21] defined several nugget-based, set retrieval metrics for the distillation task; Allan, Carterette and Lewis proposed a character-based version of *bpref* to evaluate a ranked list of passages [2]. Yang and Lad [22] have also discussed nugget-based evaluation measures that are similar in spirit to α-nDCG, for multiple queries issued over a period of time and multiple ranked lists of retrieved passages. In Yang and Lad's model, *utility* is defined as *benefit* subtracted by *cost of reading*. Whereas, in the S-measure framework, the cost of reading is used for directly discounting the value of iUnits.

2.4 Evaluating Question Answering

In Question Answering (QA), evaluation approaches similar to those for summarisation exist. *POURPRE*, an automatic evaluation metric for complex QA, is essentially F-measure computed based on unigram matches between the system output and gold-standard nuggets [10]. As in summarisation, the matching between system outputs and gold-standard nuggets can also be done manually. Either way, the main problem with this approach is that *precision* is difficult to define: while we can count the number of gold-standard nuggets present in a system output, we cannot count the number of "incorrect nuggets" in the same output. To overcome this, an *allowance* of 100 characters per nugget match

[*2] https://inex.mmci.uni-saarland.de/tracks/snippet/

[*3] http://research.microsoft.com/en-us/people/tesakai/1click2.aspx

[*4] http://www.nist.gov/tac/2011/Summarization/

was introduced at the TREC QA track; the NTCIR ACLIA task determined the allowance parameters based on average nugget lengths [11].

S-measure, in contrast, does not require the allowance parameter. While the allowance parameter implies that every nugget requires a fixed amount of space within the system output, the S-measure framework requires a *vital string* for each iUnit, based on the view that different pieces of information require different lengths of text to convey the information to the user (See Section 3.1).

One limitation of S is that it can only evaluate the *content* of the system output, just like all other nugget-based approaches. At 1CLICK-1, *readability* and *trustworthiness* ratings were obtained in parallel with the manual iUnit matches [18], which we will not discuss further in this paper.

3. NTCIR-9 1CLICK-1 Task

3.1 Task and Data

1CLICK-1, the first round of the One Click Access task, was run between March and December 2011. The task used 60 Japanese search queries, 15 for each *question category*: CELEBRITY, LOCAL, DEFINITION and QA. The CELEBRITY and LOCAL queries were selected from a mobile query log; the DEFINITION and QA queries were selected from Yahoo! Chiebukuro (Japanese Yahoo! Answers). The four query types were selected based on a query log study [8]. Two types of runs were allowed: DESKTOP runs ("D-runs") and MOBILE runs ("M-runs"), whose target lengths were $X = 500, 140$, respectively.

For a CELEBRITY query, for example, participating systems were expected to return important biography information. They were expected to return important iUnits first, and to minimise the amount of text the user has to read. For example, the iUnits for Query "Osamu Tezuka" (a famous Japanese cartoonist who died in 1989) represented his date of birth, place of birth, his occupation, the comic books he published and so on. The iUnit that represented his date of birth contained a *vital string* "1928.11.03" because this string (or something equivalent) is probably *required* in order to convey to the user that "Osamu Tezuka was born in November 3, 1928." The length of the vital string is used for defining an "optimal" output and for computing S. Moreover, at 1CLICK-1, each iUnit was weighted based on votes from five assessors.

Only three teams participated in the task, but ten runs based on diverse approaches were submitted to it: Teams KUIDL, MSRA and TTOKU took information extraction, passage retrieval and muti-document summarisation approaches, respectively[*5]. Both organisers and participants took part in manual iUnit matching, using a dedicated interface which can record match positions. Every X-string was evaluated by two assessors: in this study, we evaluate runs based on the *Intersection* data (**I**) and the *Union* data (**U**) of the iUnit matches [18]. The 60 queries and the official evaluation results are publicly available[*6], and the iUnit data can be obtained from National Institute of Informatics, Japan[*7].

For more details on 1CLICK-1, the reader is referred to the Overview paper [18]. Currently, the second round of 1CLICK (1CLICK-2) is underway.

3.2 S-measure and $S\flat$

S-measure was the primary evaluation measure used at 1CLICK-1. Let N be the set of gold-standard iUnits constructed for a particular query, and let $v(n)$ be the vital string and let $w(n)$ be the weight for iUnit $n \in N$. 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) [17]. Thus, the basic assumptions are that (a) important iUnits should be presented first; and (b) if two iUnits are equally important, then the one that can "save more space" should be presented first. The crude assumptions obviously may conflict with text readability, but have proven to be useful [17], [18]. Let $\mathit{offset}^*(v(n))$ denote the offset position of $v(n)$ within the PMO. Let $M(\subseteq N)$ denote the set of *matched* iUnits obtained by manually comparing the X-string with the gold standard iUnits, and let $\mathit{offset}(m)$ denote the offset position of $m \in M$. Morever, let L be a parameter that represents how the user's patience runs out: the original paper that proposed S used $L = 1,000$, while 1CLICK-1 used $L = 500$. 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. S is defined as:

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

Thus, all iUnits that appear after L characters within the X-string are considered worthless. When L is set to a very large value, S reduces to *weighted recall* (W-recall), which is position-insensitive. Also, as there is no theoretical guarantee that S lies below one, *S-flat* given by $S\flat = \min(1, S\text{-}\mathit{measure})$ may be used instead. In practice, the raw S values were below one for all of the submitted 1CLICK-1 runs and the "flattening" was unnecessary [18].

4. Research Questions and Proposals

4.1 Effect of the Patience Parameter

The official 1CLICK-1 evaluation used $L = 500$ (one minute) with S. In the present study, we vary this parameter as follows and examine the outcome: $L = 1,000$ (two minutes, the original setting from Sakai, Kato and Song [17]), $L = 250$ (30 seconds) and $L = 50$ (6 seconds). Note that if L is set to an extremely small value, most of the contents of the X-strings will be ignored. This is analogous to truncating ranked lists of documents prior to IR evaluation.

4.2 Evaluating Terseness: T-measure, $T\flat$ and $S\sharp$

As was discussed earlier, S cannot distinguish between Fig. 1(b) and (c). We therefore introduce a precision-like "Terseness" measure for evaluating an X-string of size X':

[*5] Note that a limited number of participants is not necessarily a weakness, as the iUnits were extracted *a priori*: the 1CLICK-1 test collection construction did not rely on pooling submitted runs [18].

[*6] http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/Evaluations/INTENT/ntc9-1CLICK-eval.htm.

[*7] http://research.nii.ac.jp/ntcir/data/data-en.html

$$T\text{-}measure = \frac{\sum_{m \in M} |v(m)|}{|X'|} . \quad (2)$$

Note that the numerator is a sum of vital string lengths, and that these lengths vary, unlike traditional nugget precision. As T might exceed one, we also define *T-flat* given by $T\flat = \min(1, T\text{-}measure)$, although in reality T never exceeded one for our data and therefore $T\flat = T$ holds. Finally, following the approach of the well-known F-measure, we can define *S-sharp* as:

$$S\sharp = \frac{(1+\beta^2)T\flat S\flat}{\beta^2 T\flat + S\flat} \quad (3)$$

where letting $\beta = 1$ reduces $S\sharp$ to a harmonic mean of S♭ and T♭. However, as we regard S as the primary measure and want T to "enter into the calculation only as a length penalty" [10], we also examined $\beta = 3, 5, 10, 20$. While $\beta = 3, 5$ reflect the practices in QA evaluation [10], [11], our experiments suggest that an even higher β may be suitable for 1CLICK, as we shall see later.

To sum up, $S\sharp$ differs from the traditional nugget-based F-meaure in the following two aspects: (1) It utilises the positions of iUnits for computing the recall-like S; and (2) Instead of relying on a fixed allowance parameter, it utilises the vital string length of each iUnit for computing the precision-like T-measure.

5. Experiments

5.1 Results on the Patience Parameter

Figure 2(a) and (b) show the effect of L on the overall system ranking with Mean S with **I** and with **U**, respectively. The x-axis shows the runs sorted by Mean S ($L = 500$), i.e. the official ranking. With **I**, Kendall's τ with the official ranking are .87 (Mean W-recall), .96 ($L = 1,000$), .78 ($L = 250$) and .64 ($L = 50$); with **U**, the corresponding values are .82 (Mean W-recall), .96 ($L = 1,000$), .73 ($L = 250$) and .69 ($L = 50$). Thus, $L = 1,000$ (two minutes [17]) produces rankings that are very similar to $L = 500$ (one minute), but $L = 250$ (30 seconds) results in substantially different system rankings. In particular, Fig. 2(a) shows that while Mean S with $L = 500$ prefers KUIDL-D-OPEN-1 over MSRA1click-D-OPEN-2 and prefers KUIDL-D-OPEN-2 over MSRA1click-D-OPEN-1, Mean S with $L = 250$ has exactly the opposite preferences. This trend is further emphasized by Mean S with $L = 50$.

Recall that S with $L = 250$ *ignores* all iUnit matches between positions 250 and 500 for all of the D-runs. Thus, the above discrepancy between $L = 500$ and $L = 250$ regarding KUIDL and MSRA1click suggests that *while* KUIDL *is good at covering important iUnits,* MSRA1click *is good at presenting the most important units near the beginning of the X-string.* To illustrate this point, **Fig. 3** shows the actual X-strings of KUIDL and MSRA1click for a LOCAL query "*Menard Aoyama Resort*" (name of a facility). It can be observed that even though KUIDL is superior to MSRA1click in terms of the number of matches with **I** (4 matches vs. 3), MSRA1click is actually very good from the viewpoint of iUnit *positions* as indicated by the underlined texts that correspond to the iUnit matches. With **I**, the S with $L = 500$ for KUIDL is 0.200, and that for MSRA1click is 0.332; whereas, the S with $L = 250$ for KUIDL is 0.120, and that for MSRA1click is 0.528. Thus the difference between two systems is magnified when $L = 250$.

Next, we examine the effect of L on *discriminative power*. Given a test collection with a set of runs, discriminative power is measured by conducting a statistical significance test for every pair of runs [15]. This methodology has been used in a number of evaluation studies [5], [14], [16], [19], [20], and is arguably one necessary (but by no means sufficient) condition of a "good" measure. We used a *randomised version of Tukey's Honestly Significant Differences (HSD) test* for testing statistical significance, which is known to be more reliable than traditional *pairwise* significance tests [4], [16].

Figure 4 shows the *Achieved Siginificance Level (ASL) curves* [15] of S with varying L. Here, the y-axis represents the ASL (i.e. p-value), and the x-axis represents the 45 run pairs sorted by the p-value. Measures that are closer to the origin are the ones that are highly discriminative, i.e. those that provide reliable experimental results. It can be observed that the discriminative power for $L = 250$ is the highest while that for $L = 50$ is low (naturally, as the latter implies looking at only the first 50 characters of every X-string). Moreover, S with $L = 250$ is more discriminative than W-recall. These observations are consistent across **I** and **U**. Thus, at least for the runs submitted to 1CLICK-1, using $L = 250$ (user has 30 seconds) along with the official $L = 500$ (user has one minute) seems beneficial not only for examining 1CLICK systems from different angles but also for enhancing discriminative power. Based on these results, we consider $L = 250, 500$ in the next section.

5.2 Results on T-measure and $S\sharp$

Next, we discuss T and $S\sharp$, which we introduced for penalising redundancy in 1CLICK evaluation. **Figure 5** shows the system rankings according to Mean S, T and $S\sharp$ (where the x-axis represents runs sorted by Mean S with $L = 500$), while **Fig. 6** shows the kendall's τ between the ranking by Mean S and one by Mean $S\sharp$ with β (denoted by $S\sharp\beta$). Note that β means "S is β times as important as T" and that $S\sharp 0 = T$ (See Eq. 3).

Figure 5 shows that T rates the four M-runs that contain "-M-" in their run names (especially the two KUIDL-M runs) relatively highly, but this is because M-runs use $X = 140$ as the target length while D-runs use $X = 500$. (Had the 1CLICK-1 task received more runs, these two run types would have been ranked separately.) More interestingly, The Mean $S\sharp$ rankings in Fig. 5(a) unanimously prefer MSRA1click-D-OPEN-2 over KUIDL-D-OPEN-1 and prefer MSRA1click-D-OPEN-1 over KUIDL-D-OPEN-2, contrary to the official Mean S ranking. This suggests that MSRA1click was actually better than KUIDL from the viewpoint of terseness. To illustrate this point, **Fig. 7** shows the X-strings for the QA query "*The three duties of a Japanese citizen*": both KUIDL and MSRA1click managed to capture the three answers and their S values are 0.977 and 0.988, respectively (note that the former underperforms the latter even in terms of S, due to one ill-placed iUnit); whereas, the T values are 0.014 and 0.400, respectively. Thus, T reflects the fact that the X-string of KUIDL is highly redundant while that of MSRA1click is almost perfect. (The figure shows how to compute S and T for the X-string of MSRA1click.) It can be observed that T and $S\sharp$ are

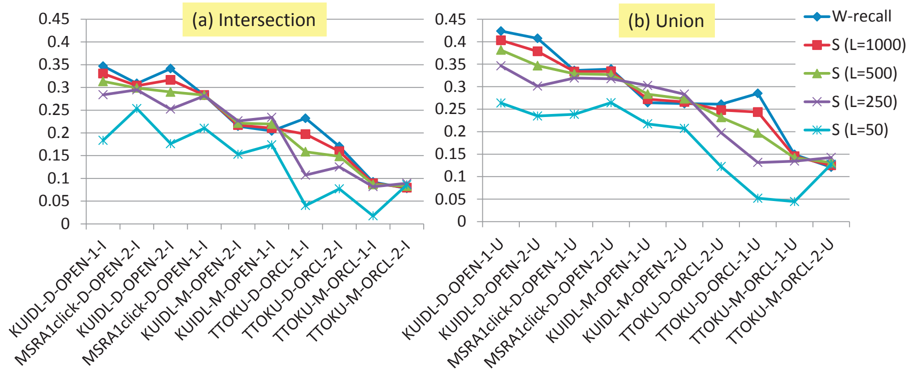
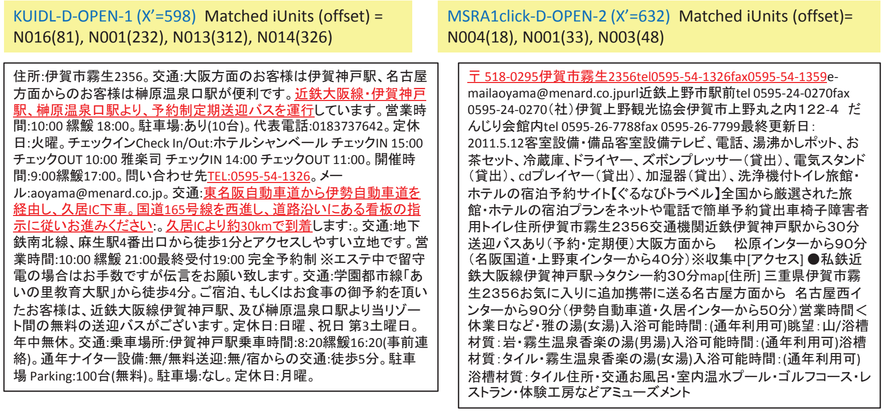


Fig. 2 Effect of the patience parameter L on the system ranking. The x-axis shows runs sorted by Mean S-measure ($L = 500$), i.e. the official ranking.


KUIDL-D-OPEN-1 (X'=598) Matched iUnits (offset) = N016(81), N001(232), N013(312), N014(326)

住所:伊賀市霧生2356。交通:大阪方面のお客様は伊賀神戸駅、名古屋方面からのお客様は榊原温泉口駅が便利です。近鉄大阪線・伊賀神戸駅、榊原温泉口駅より、予約制定期送迎バスを運行しています。営業時間:10:00 縲鰀 18:00。駐車場:あり(10台)。代表電話:0183737642。定休日:火曜。チェックインCheck In/Out:ホテルシャンベール チェックIN 15:00 チェックOUT 10:00 雅楽司 チェックIN 14:00 チェックOUT 11:00。開催時間:9:00縲鰀17:00。問い合わせ先TEL:0595-54-1326。メール:aoyama@menard.co.jp。交通:東名阪自動車道から伊勢自動車道を経由し、久居IC下車。国道165号線を西進し、道路沿いにある看板の指示に従いお進みください:。久居ICより約30kmで到着します:。交通:地下鉄南北線、麻生駅4番出口から徒歩1分とアクセスしやすい立地です。営業時間:10:00 縲鰀 21:00最終受付19:00 完全予約制 ※エステ中で留守電の場合はお手数ですが伝言をお願い致します。交通:学園都市線「あいの里教育大駅」から徒歩4分。ご宿泊、もしくはお食事の御予約を頂いたお客様は、近鉄大阪線伊賀神戸駅、及び榊原温泉口駅より当リゾート間の無料の送迎バスがございます。定休日:日曜 、祝日 第3土曜日。年中無休。交通:乗車場所:伊賀神戸駅乗車時間:8:20縲鰀16:20(事前連絡)。通年ナイター設備:無/無料送迎:無/宿からの交通:徒歩5分。駐車場 Parking:100台(無料)。駐車場:なし。定休日:月曜。

MSRA1click-D-OPEN-2 (X'=632) Matched iUnits (offset)= N004(18), N001(33), N003(48)

〒 518-0295伊賀市霧生2356tel0595-54-1326fax0595-54-1359e-mailaoyama@menard.co.jpurl近鉄上野市駅前tel 0595-24-0270fax 0595-24-0270（社）伊賀上野観光協会伊賀市上野丸之内122-4 だんじり会館内tel 0595-26-7788fax 0595-26-7799最終更新日：2011.5.12客室設備・備品客室設備テレビ、電話、湯沸かしポット、お茶セット、冷蔵庫、ドライヤー、ズボンプレッサー（貸出）、電気スタンド（貸出）、cdプレイヤー（貸出）、加湿器（貸出）、洗浄機付トイレ旅館・ホテルの宿泊予約サイト【ぐるなびトラベル】全国から厳選された旅館・ホテルの宿泊プランをネットや電話で簡単予約貸出車椅子障害者用トイレ住所伊賀市霧生2356交通機関近鉄伊賀神戸駅から30分送迎バスあり（予約・定期便）大阪方面から　松原インターから90分（名阪国道・上野東インターから40分）※収集中[アクセス] ●私鉄近鉄大阪線伊賀神戸駅→タクシー約30分map[住所] 三重県伊賀市霧生2356お気に入りに追加携帯に送る名古屋方面から　名古屋西インターから90分（伊勢自動車道・久居インターから50分）営業時間＜休業日など・雅の湯(女湯)入浴可能時間：(通年利用可)眺望：山/浴槽材質：岩・霧生温泉香楽の湯(男湯)入浴可能時間：(通年利用可)浴槽材質：タイル・霧生温泉香楽の湯(女湯)入浴可能時間：(通年利用可)浴槽材質：タイル住所・交通お風呂・室内温水プール・ゴルフコース・レストラン・体験工房などアミューズメント

Fig. 3 X-strings of runs from KUIDL and MSRA1click for the LOCAL query "*Menard Aoyama Resort*" (name of a facility).

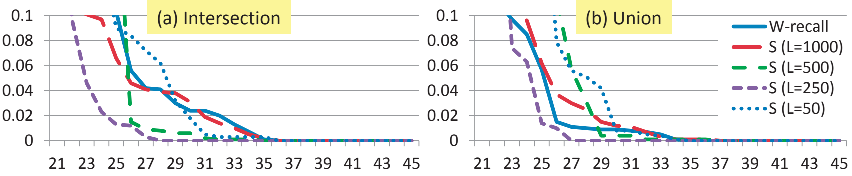


Fig. 4 Effect of the patience parameter L on discriminative power. The y-axis represents the p-value and the x-axis represents run pairs sorted by the p-value.

useful complements to S for evaluating 1CLICK systems.

Figure 8 shows the ASL curves for our proposed measures. From the viewpoint of discriminative power, it can be observed that T is very poor, and therefore that it is safer to set β to a high value when using $S\sharp$. To be more specific, it can be observed that the discriminative power of $S\sharp10$ is comparable to that of S for both $L = 250$ (shown as dotted lines) and $L = 500$ (shown as solid lines). Since $S\sharp10$ retains the high discriminative power of S *and* provide new insight to the evaluation as shown in Fig. 5 and Fig. 6, we recommend $S\sharp10$ for evaluating 1CLICK systems, along with the original S.

6. Conclusions and Future Work

This paper extended the 1CLICK-1 evaluation framework, where systems were required to return a single, concise textual output in response to a query in order to satisfy the user immediately after a click on the SEARCH button. We first showed that the discount parameter L of S-measure affects system ranking and discriminative power, and that using multiple values, e.g. $L = 250$ (user has only 30 seconds to view the text) and $L = 500$ (user has one minute), is useful: a 1CLICK system which can satisfy the user's information need within one minute may be different from one which can satisfy the need within 30 seconds. Also, S with $L = 250$ appears to be more discriminative than S with $L = 500$ and W-recall, at least for the runs submitted to the 1CLICK-1 task.

We then complemented the recall-like S with a simple, precision-like measure called T-measure as well as a combination of S and T, called $S\sharp$. We showed that $S\sharp$ with a heavy emphasis on S (e.g. $S\sharp10$) imposes an appropriate length penalty to 1CLICK-1 system outputs and yet achieves discriminative power that is comparable to S. These new measures will be used at the NTCIR-10 1CLICK-2 task, where we hope to experiment with more participating teams and runs.

At 1CLICK-2, the language scope has been extended to English and Japanese. While the evaluation framework of S, T and $S\sharp$ should apply to any language, it would be interesting to test it in the English subtask as well. There may be language-dependent issues in defining iUnits and vital strings[*8]. More-

[*8] The new definitions of iUnits and vital strings for the Japanese 1CLICK-

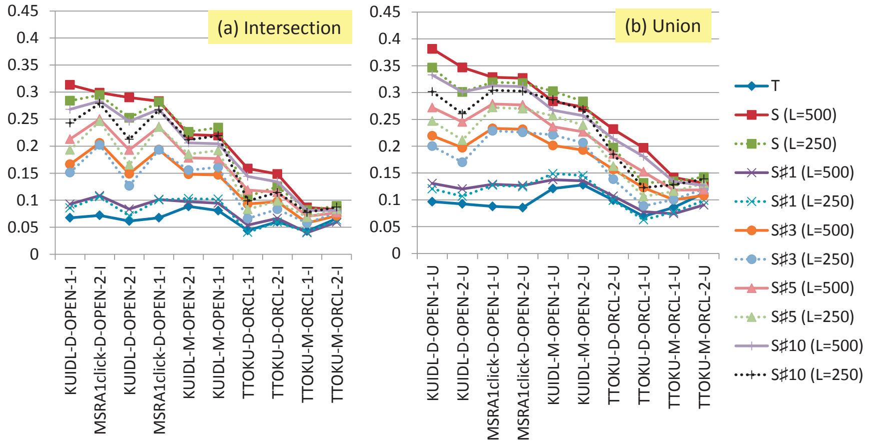


Fig. 5 System ranking by different measures. The *x*-axis shows runs sorted by Mean S-measure with $L = 500$, i.e., the official ranking.

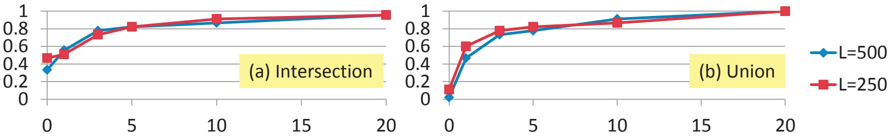
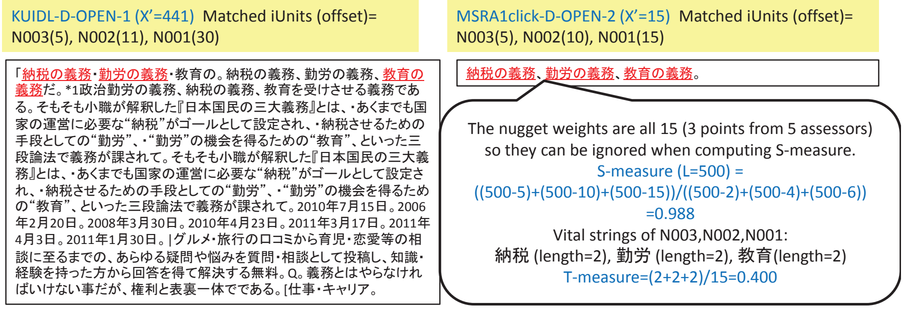


Fig. 6 Effect of β on $S\sharp\beta$: the *x*-axis represents β and the *y*-axis represents Kendall's τ with the Mean S-measure ranking.



KUIDL-D-OPEN-1 (X'=441) Matched iUnits (offset)= N003(5), N002(11), N001(30)

「納税の義務・勤労の義務・教育の。納税の義務、勤労の義務、教育の義務だ。*1政治勤労の義務、納税の義務、教育を受けさせる義務である。そもそも小職が解釈した『日本国民の三大義務』とは、・あくまでも国家の運営に必要な“納税”がゴールとして設定され、・納税させるための手段としての“勤労”、・“勤労”の機会を得るための“教育”、といった三段論法で義務が課されて。そもそも小職が解釈した『日本国民の三大義務』とは、・あくまでも国家の運営に必要な“納税”がゴールとして設定され、・納税させるための手段としての“勤労”、・“勤労”の機会を得るための“教育”、といった三段論法で義務が課されて。2010年7月15日。2006年2月20日。2008年3月30日。2010年4月23日。2011年3月17日。2011年4月3日。2011年1月30日。|グルメ・旅行の口コミから育児・恋愛等の相談に至るまでの、あらゆる疑問や悩みを質問・相談として投稿し、知識・経験を持った方から回答を得て解決する無料。Q。義務とはやらなければいけない事だが、権利と表裏一体でである。[仕事・キャリア。

MSRA1click-D-OPEN-2 (X'=15) Matched iUnits (offset)= N003(5), N002(10), N001(15)

納税の義務、勤労の義務、教育の義務。

The nugget weights are all 15 (3 points from 5 assessors) so they can be ignored when computing S-measure.
S-measure (L=500) = ((500-5)+(500-10)+(500-15))/((500-2)+(500-4)+(500-6)) =0.988
Vital strings of N003,N002,N001:
納税 (length=2), 勤労 (length=2), 教育(length=2)
T-measure=(2+2+2)/15=0.400

Fig. 7 *X*-strings of runs from KUIDL and MSRA1click for the QA query "*The three duties of a Japanese citizen.*"

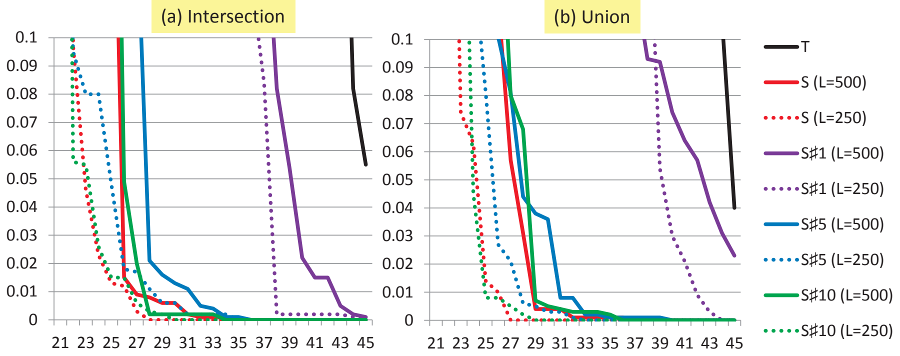


Fig. 8 Discriminative power of S-measure, T-measure and $S\sharp$.

over, we plan to look into the relationship between these measures with readability, trustworthiness and other qualities required of an *X*-string [18], and the relationship between these measures with measures based on *automatic* matching [13].

[2] subtask can be found at `http://www.dl.kuis.kyoto-u.ac.jp/~kato/1click2/data/1C2-J-SAMPLE-README.pdf`.

References

[1] Allan, J., Aslam, J., Azzopardi, L., Belkin, N., Borlund, P., Bruza, P., Callan, J., Carman, M., Clarke, C.L., Craswell, N., Croft, W.B., Culpepper, J.S., Diaz, F., Dumais, S., Ferro, N., Geva, S., Gonzalo, J., Hawking, D., Jarvelin, K., Jones, G., Jones, R., Kamps, J., Kando, N., Kanoulas, E., Karlgren, J., Kelly, D., Lease, M., Lin, J., Mizzaro, S., Moffat, A., Murdock, V., Oard, D.W., de Rijke, M., Sakai, T., Sanderson, M., Scholer, F., Si, L., Thom, J.A., Thomas, P., Trotman, A., Turpin, A., de Vries, A.P., Webber, W., Zhang, X., , Zhang, Y.: Frontiers, challenges and opportunities for information retrieval: Report from SWIRL 2012. SIGIR Forum 46(1), 2–32 (2012)

[2] Allan, J., Carterette, B., Lewis, J.: When will information retrieval be

"good enough"? In: Proceedings of ACM SIGIR 2005. pp. 433–440 (2005)

[3] Babko-Malaya, O.: Annotation of nuggets and relevance in gale distillation evaluation. In: Proceedings of LREC 2008. pp. 3578–3584 (2008)

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

[5] Clarke, C.L., Craswell, N., Soboroff, I., Ashkan, A.: A comparative analysis of cascade measures for novelty and diversity. In: Proceedings of ACM WSDM 2011 (2011)

[6] Dunlop, M.D.: Time, relevance and interaction modelling for information retrieval. In: Proceedings of ACM SIGIR '97 (1997)

[7] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20(4), 422–446 (2002)

[8] Li, J., Huffman, S., Tokuda, A.: Good abandonment in mobile and PC internet search. In: Proceedings of ACM SIGIR 2009. pp. 43–50 (2009)

[9] Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the ACL 2004 Workshop on Text Summarization Branches Out (2004)

[10] Lin, J., Demner-Fushman, D.: Methods for automatically evaluating answers to complex questions. Information Retrieval 9(5), 565–587 (2006)

[11] Mitamura, T., Shima, H., Sakai, T., Kando, N., Mori, T., Takeda, K., Lin, C.Y., Song, R., Lin, C.J., Lee, C.W.: Overview of the NTCIR-8 ACLIA tasks: Advanced cross-lingual information access. In: Proceedings of NTCIR-8. pp. 15–24 (2010)

[12] Nenkova, A., Passonneau, R., McKeown, K.: The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Transactions on Speech and Language Processing 4(2), Article 4 (2007)

[13] Pavlu, V., Shahzad, Rajput, Golbus, P.B., Aslam, J.A.: IR system evaluation using nugget-based test collections. In: Proceedings of ACM WSDM 2012 (2012)

[14] Robertson, S.E., Kanoulas, E., Yilmaz, E.: Extending average precision to graded relevance judgments. In: Proceedings of ACM SIGIR 2010. pp. 603–610 (2010)

[15] Sakai, T.: Evaluating evaluation metrics based on the bootstrap. In: Proceedings of ACM SIGIR 2006. pp. 525–532 (2006)

[16] Sakai, T.: Evaluation with informational and navigational intents. In: Proceedings of WWW 2012. pp. 499–508 (2012)

[17] Sakai, T., Kato, M.P., Song, Y.I.: Click the search button and be happy: Evaluating direct and immediate information access. In: Proceedings of ACM CIKM 2011. pp. 621–630 (2011)

[18] Sakai, T., Kato, M.P., Song, Y.I.: Overview of NTCIR-9 1CLICK. In: Proceedings of NTCIR-9. pp. 180–201 (2011)

[19] Soboroff, I.: Test collection diagnosis and treatment. In: Proceedings of EVIA 2010. pp. 34–41 (2010)

[20] Webber, W., Moffat, A., Zobel, J.: The effect of pooling and evaluation depth on metric stability. In: Proceedings of EVIA 2010. pp. 7–15 (2010)

[21] White, J.V., Hunter, D., Goldstein, J.D.: Statistical evaluation of information distillation systems. In: Proceedings of LREC 2008. pp. 3598–3604 (2008)

[22] Yang, Y., Lad, A.: Modeling expected utility of multi-session information distillation. In: Proceedings of ICTIR 2009. pp. 164–175 (2009)