Chapter 11

Evaluation of Information Access with Smartphones

Makoto P. Kato

Abstract NTCIR 1CLICK and MobileClick are the earliest attempts toward test-collection-based evaluation for information access with smartphones. Those campaigns aimed to develop an IR system that outputs a short text summary for a given query, which is expected to fit a small screen and to satisfy users' information needs without requiring much interaction. The textual output was evaluated on the basis of *iUnits*, pieces of relevant text for a given query, with consideration of users' reading behaviors. This chapter begins with an introduction to NTCIR 1CLICK and MobileClick, explains the evaluation methodology and metrics such as S-measure and M-measure, and finally discusses the potential impacts of those evaluation campaigns.

11.1 Introduction

In 2015, Google announced that more searches took place on mobile devices than on desktop computers in 10 countries including the US and Japan.¹ Among diverse types of mobile devices, the smartphone has become dominant according to a survey in 2015.² Thus, there is no doubt that the smartphone is one of the most important search environments for which search engines should be designed, due to its popularity and several differences from traditional devices, e.g., desktop computers.

The search experience difference between desktop computers and smartphones mainly comes from the differences in screen size, internet connection, interaction, and situation. A relatively small screen size limits the amount of content which the users can read at a time. The internet connection is sometimes unstable depending on

¹<https://adwords.googleblog.com/2015/05/building-for-next-moment.html>.

²<https://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/>.

M. P. Kato (✉)

University of Tsukuba, 1-2 Kasuga, Tsukuba, Ibaraki 305-8550, Japan

e-mail: mpkato@acm.org

where users conduct search. While the keyboard and mouse are typical input devices for desktop computers, touch interaction and speech input are often used for smartphones and may not be suitable for inputting or editing many keywords. Search with smartphones can sometimes be interrupted by the other activities with which the user is engaged simultaneously. To overcome the limitations in search with smartphones, research communities have studied new designs of interface, interaction, and search algorithms suitable for smartphones (Crestani et al. 2017).

NTCIR 1CLICK (Kato et al. 2013a; Sakai et al. 2011b) and MobileClick (Kato et al. 2014, 2016b) are the earliest attempts toward test-collection-based evaluation for information access with smartphones. Those campaigns aimed to develop an IR system that outputs a short text summary for a given query, which is expected to fit a small screen and to satisfy users' information needs without requiring much interaction. The textual output was evaluated on the basis of pieces of relevant text for a given query. The basic task design is similar to *query-biased multi-document summarization* (Carbonell and Goldstein 1998; Tombros and Sanderson 1998), in which a system is expected to generate a summary of a fixed length from multiple documents, satisfying the information need of users who input a certain query. The main difference from the query-biased multi-document summarization task is *position awareness* of presented information. In the NTCIR 1CLICK and MobileClick tasks, more important information is expected to be present at the beginning of the summary so that users can reach such information efficiently. In other words, more relevant information pieces should be ranked at higher positions like an ad hoc retrieval task. Accordingly, evaluation measures used in these tasks were designed to be position-aware, unlike those for text summarization such as recall, precision, and ROUGE (Lin 2004). This task design and evaluation methodology distinguishes NTCIR 1CLICK and MobileClick from the other summarization tasks, and had some impact on mobile information access and related fields.

This chapter first describes the task design of NTCIR 1CLICK and MobileClick, introduces evaluation methodologies used in these campaigns, and finally discusses potential impacts on works published after NTCIR 1CLICK and MobileClick.

11.2 NTCIR Tasks for Information Access with Smartphones

This section provides a brief overview of the task design of the NTCIR 1CLICK and MobileClick tasks. Table 11.1 summarizes four NTCIR tasks to be described in this section.³

³S_#-measure is a combination of S-measure and T-measure (a precision-like metric) (Sakai and Kato 2012).

Table 11.1 NTCIR tasks for information access with smartphones

Year	NTCIR	Task	Subtasks	Primary metric
2010	9	1CLICK-1		S-measure
2011	10	1CLICK-2	Main & Query classification	S _q -measure ³
2013	11	MobileClick-1	iUnit retrieval & iUnit summarization	M-measure
2014	12	MobileClick-2	iUnit retrieval & iUnit summarization	M-measure

11.2.1 NTCIR 1CLICK

The history of information access with smartphones in NTCIR began from a subtask of the NTCIR-9 INTENT task, namely, NTCIR-9 1CLICK-1 (formally, *one-click access task*) (Sakai et al. 2011b). While the NTCIR-9 INTENT task targeted search result diversification, the NTCIR-9 1CLICK-1 task focused especially on generating a query-biased summary as a proxy for a search engine result page (or “ten blue links”), for satisfying the user immediately after the user clicks on the search button. Strictly speaking, the NTCIR-9 1CLICK-1 task was designed not for information access with smartphones, but for *Direct and Immediate Information Access*, which was defined in earlier work by the task organizers (Sakai et al. 2011a):

We define Direct Information Access as a type of information access where there is no user operation such as clicking or scrolling between the user’s click on the search button and the user’s information acquisition; we define Immediate Information Access as a type of information access where the user can locate the relevant information within the system output very quickly. Hence, a Direct and Immediate Information Access (DIIA) system is expected to satisfy the user’s information need very quickly with its very first response.

While the NTCIR-9 1CLICK-1 task was treated as a pilot task and targeted only the Japanese language, the 1CLICK-2 task was organized as an independent task at NTCIR-10 and employed almost the same task design as that of the NTCIR-9 1CLICK-1 task, with the scope extended to Japanese and English.

At both 1CLICK-1 and 1CLICK-2, participants were given a list of queries categorized into four query categories, namely, celebrity, local, definition, and Q&A. The task organizers selected these categories the following work by Li et al. (2009), which investigated Google’s desktop and mobile query logs of three countries, and identified frequent query types for *good abandonment*—an abandoned query for which the user’s information need was successfully addressed by the search engine result page without clicks or query reformulation.

NTCIR-9 1CLICK-1 and NTCIR-10 1CLICK-2 participants were expected to produce a plain text of X characters for each query ($X = 140$ for Japanese and $X =$

Fig. 11.1 A two-layered summary for query “christopher nolan”. Users can see the second layer if they click on a link in the first layer

280 for English),⁴ based on a given document collection. The output was expected to include important pieces of information first and to minimize the amount of text the user has to read. These requirements are more formally described through the evaluation metrics explained in Sect. 11.3.

11.2.2 NTCIR MobileClick

NTCIR MobileClick, which started from NTCIR-11, took over the spirit of NTCIR 1CLICK, and aimed to directly return a summary of relevant information and immediately satisfy the user *without requiring much interaction*. Unlike the 1CLICK tasks, participants were expected to produce a two-layered summary that consists of a single first layer and multiple second layers, as shown in Fig. 11.1. The first layer is expected to contain information interesting for most of the users, and the links to the second layer; the second layer, which is hidden until its header link is clicked on, is expected to contain information relevant for a particular type of users. In a two-layered summary, users can avoid reading text in which they are not interested, thus saving time spent on non-relevant information, if they can make a binary yes/no decision of each second-layer entry from the head link alone.

⁴Both NTCIR-9 1CLICK-1 and NTCIR-10 1CLICK-2 accepted two types of runs, namely, DESKTOP and MOBILE runs. In this chapter, only MOBILE runs are explained for simplicity.

This unique output was motivated by the discussion at the NTCIR-10 conference in June 2013, and reflected the rapid growth of smartphone users in those years. Although 1CLICK expects no interaction except for clicking on the search button, MobileClick targeted smartphone users and expects users to tap on some links for browsing desired information efficiently.

NTCIR MobileClick assumed different types of users who are interested in different topics. The diversity of users who input a certain query was modeled by *intent probability*, which is the probability over intents for the query. For example, among users who input “apple” as a query, 90% are interested in Apple Inc. and 10% are interested in apple the fruit. A two-layered summary is considered *good* if different types of users are all satisfied with the summary. Thus, the first layer should not contain information in which a particular type of users are interested, while the second layers should not contain information relevant to the majority of users.

The input in the NTCIR-11 MobileClick-1 and NTCIR-12 MobileClick-2 tasks was a list of queries that were basically categorized into four types mentioned earlier. There were two subtasks in these evaluation campaigns: iUnit retrieval and iUnit summarization subtasks. In iUnit retrieval subtask, participants were expected to output a ranked list of information pieces called iUnit in response to a given query. In iUnit summarization subtask, as was explained earlier, the output was a two-layered summary in XML format. While the NTCIR-11 MobileClick-1 required participants to identify information pieces from a document collection, the NTCIR-12 MobileClick-2 only required selecting and ranking or arranging predefined information pieces, mainly for increasing the reusability of the test collection.

11.3 Evaluation Methodology in NTCIR 1CLICK and MobileClick

This section explains and discusses some details of the evaluation methodology used in the NTCIR 1CLICK and MobileClick tasks, which is mainly based on nuggets, or pieces of information we call *iUnits*. We first present the background and explain the differences between summarization and our tasks. We then focus on the notions of nuggets and iUnits, and finally discuss the effectiveness metrics developed and used in the NTCIR tasks.

11.3.1 Textual Output Evaluation

Summarization is one of the most similar tasks to NTCIR 1CLICK and MobileClick. As mentioned earlier, the most notable difference between the summarization and these NTCIR tasks is position awareness of information pieces in the textual output. This subsection details and discusses the difference in terms of the evaluation methodology.

Automatic evaluation of machine-generated summaries has been often conducted by comparison with human-generated summaries (Nenkova and McKeown 2011). ROUGE is a widely used evaluation metric based on word matching between a machine summary and human summaries (Lin 2004). There are several variants of ROUGE such as ROUGE-W (n -gram matching), ROUGE-L (longest common sequence), and ROUGE-S (skip-gram matching). Although these variants are sensitive to the order of words, they are agnostic to the absolute position where each word appears in a machine summary. The Pyramid method identifies Summary Content Units (SCUs), which are word spans expressing the same meaning, from multiple human summaries, and computes a score for each machine summary based on the included SCUs (Nenkova et al. 2007). The weight of an SCU is determined by the number of human summaries including the SCU, and a summary is scored basically by the sum of the weights of SCUs within the summary. The position of SCUs within a machine summary does not affect the score.

The insensitivity for the position of information pieces (*i.e.*, words or SCUs) is reasonable when it is assumed that the whole summary is always read. In such a case, the position of information pieces should not affect the utility of the summary, as all the information pieces are equally consumed by the reader.

On the other hand, the position matters when users may read different parts of a summary. As the textual output in NTCIR 1CLICK is expected to be scanned from top to bottom, like Web search, contents near the end have a smaller probability to be read, and, accordingly, should be discounted when the utility is estimated. The two-layer summary in NTCIR MobileClick can be read in many different ways. A user may read only the first layer, while another user may scan contents in the first layer from top, click on a link interesting for the user, read a second layer shown by the click, and stop reading at the end of the second layer. Therefore, the primary difference from ordinary summarization tasks is how the summary is expected to be read, which naturally required different evaluation methodologies.

11.3.2 From Nuggets to *iUnits*

The NTCIR-9 1CLICK-1 task evaluated the system output based on nuggets. Nuggets are fragments of text, which were frequently used in summarization and question answering evaluation. TREC Question Answering track defined an *information nugget* as “a fact for which the assessor could make a binary decision as to whether a response contained the nugget” (Voorhees 2003). The possibility of the binary decision is called *atomicity* (Dang et al. 2007). As explained earlier, the Pyramid method (Nenkova et al. 2007) uses SCUs as units of comparison:

SCUs are semantically motivated, subsentential units; they are variable in length but not bigger than a sentential clause. This variability is intentional since the same information may be conveyed in a single word or a longer phrase. SCUs emerge from annotation of a

collection of human summaries for the same input. They are identified by noting information that is repeated across summaries, whether the repetition is as small as a modifier of a noun phrase or as large as a clause.

Babko-Malaya described a systematic way to uniform the granularity of nuggets based on several nuggetization rules (Babko-Malaya 2008). Examples of the rules are shown below:

Nuggets are created out of each core verb and its arguments, where the maximal extent of the argument is always selected.

Noun phrases are not decomposed into separate nuggets, unless they contain temporal, locative, numerical information, or titles.

Basic elements are another attempt to systematically define nuggets (Hovy et al. 2006), and were defined as follows:

the head of a major syntactic constituent (noun, verb, adjective or adverbial phrases), expressed as a single item, or a relation between a head-BE and a single dependent, expressed as a triple (head—modifier—relation).

Although several attempts had been made to standardize the nuggetization procedure, the task organizers of NTCIR 1CLICK still found it hard to identify nuggets. The primary difficulty is to uniform the granularity of nuggets. While the notion of atomicity determines the unit of nuggets to some extent, there were some cases in which assessors disagreed. Typical examples are shown below:

1. Tetsuya Sakai was born in 1988.
2. Takehiro Yamamoto received a PhD from Kyoto University in 2011.

The following pieces are candidates for nuggets in sentences 1 and 2.

- 1-A. Tetsuya Sakai was born in 1988.
- 1-B. Tetsuya Sakai was born.
- 2-A. Takehiro Yamamoto received a PhD from Kyoto University in 2011.
- 2-B. Takehiro Yamamoto received a PhD in 2011.
- 2-C. Takehiro Yamamoto received a PhD from Kyoto University.
- 2-D. Takehiro Yamamoto received a PhD.

Although 1-B and 2-D are results of a similar type of decomposition, 1-B does not look appropriate for a nugget, but 2-D does. Whereas, 2-D may not be an appropriate nugget if the query is “When did Takehiro Yamamoto receive his PhD?” since 2-D can be a trivial fact like 1-B. A systematic approach may not be very helpful in this case.

Another difficulty is the way to determine the weight of nuggets. Unlike the Pyramid method and others, the NTCIR-9 1CLICK-1 task extracted nuggets from a document collection from which the textual output is generated, not from those generated by human assessors. This methodology was chosen because there were hundreds of nuggets for some queries, which cannot be included in a few human-generated summaries. The weighting schema used in the Pyramid method cannot

be simply applied to this case, as the number of assessors who found a nugget may simply reflect the frequency of the nugget in the collection, but it might be unrelated to the importance of the nugget. Furthermore, the dependency of nuggets makes the problem more complicated. For example, 2-B entails 2-D. Then, what is the score of a summary including 2-B? Is it the sum of the weights of 2-B and 2-D, or 2-B's alone?

To clarify the definition of nuggets and weighting schema in NTCIR 1CLICK, the task organizers of the NTCIR-10 1CLICK-2 opted to redefine nuggets and call them *information units* or *iUnits*.

iUnits satisfy three properties, *relevant*, *atomic*, and *dependent*, described in detail below. *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. Note that this property emphasizes that the information need determines which pieces of information are iUnits. If the information need is “Where was Tetsuya Sakai born?”, both cannot be iUnits.

Atomic means that an 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 an 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.

In the NTCIR-10 1CLICK-2, nuggets were first identified from a document collection, and iUnits were extracted from the nuggets.⁵ A set of iUnits for query 1C2-J-0001 “倉木麻衣 (Mai Kuraki; a Japanese singer-songwriter)” is shown in Table 11.3,

⁵This approach was taken mainly for increasing the efficiency by dividing the iUnit extraction task into two parts.

Table 11.2 Nuggets for query 1C2-J-0001 “倉木麻衣 (Mai Kuraki; a Japanese singer-songwriter)”

ID	Nugget
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。(She made her American debut with “Baby I Like” as “Mai K” in October 1999, when she was 16 years old. In the same year, on December 8, she made her debut in Japan with “Love, Day After Tomorrow”.)
S008	血液型 B型 (Blood type: B)
S012	職業 歌手 (Occupation: Singer)
S022	2005年立命館大学を卒業 (She graduated from Ritsumeikan University in 2005.)
S023	第15回日本ゴールドディスク大賞で『delicious way』が ³ ロック・アルバム・オブ・ザ・イヤーを、「Secret of my heart」が ³ ソング・オブ・ザ・イヤーを受賞。(“delicious way” won “Rock album of the Year” and “Secret of my heart” won “Song of the Year” at the 15th annual Japan Gold Disc Awards)

which were extracted from nuggets in Table 11.2. The column “Entails” 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. A *semantics* is the factual statement that the iUnit conveys. This is used by assessors to determine whether an iUnit is present in a summary.

A *vital string* is a minimally adequate natural language expression and extracted from iUnits. 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 u that entails iUnits $e(u)$ does not include that of iUnits $e(u)$ to avoid duplication of vital strings, since if iUnit u is present in a summary, iUnits $e(u)$ are also present by definition. For example, the vital string of iUnit I014 does not include that of iUnit I013 as shown in Table 11.3. Even the vital string of I085 is empty as it entails iUnits I023 and I033.

Having extracted iUnits from nuggets, assessors gave the weight 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 weight so that iUnit u entailing iUnits $e(u)$ receives the weight of only u excluding that of $e(u)$. This revision is necessary because the presence of iUnit u in a summary entails that of iUnits $e(u)$, resulting in duplicative counting of the weight of $e(u)$ when we take into account the weight of both u and $e(u)$.

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).

Table 11.3 iUnits for query 1C2-J-0001 “倉木麻衣 (Mai Kuraki; a Japanese singer-songwriter)”

ID	Entails	Nugget	Semantics	Vital string
I011		S005	1999年日本デビュー (Made her Japanese debut in 1999)	1999年日本デビュー (Made her Japanese debut in 1999)
I012		S008	血液型B型 (Blood type: B)	血液型B型 (Blood type: B)
I013		S022	立命館大卒 (Graduated from Ritsumeikan University)	立命館大卒 (Graduated from Ritsumeikan University)
I014	I013	S022	2005年立命館大卒 (Graduated from Ritsumeikan University in 2005)	2005年 (2005)
I017		S012	職業 歌手 (Occupation: Singer)	歌手 (Singer)
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－受賞 (Won “Song of the Year” at the 15th annual Japan Gold Disc Awards)	第15回日本ゴールドディスク大賞ソング・オブ・ザ・イヤ－受賞 (Won “Song of the Year” at the 15th annual Japan Gold Disc Awards)
I033		S023	シングル Secret of my heart (Single “Secret of my heart”)	シングル Secret of my heart (Single “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－を受賞 (“Secret of my heart” won “Song of the Year” at the 15th annual Japan Gold Disc Awards)	

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

More formally, we used the following equation for revising the weight of iUnit u :

$$w(u) - \max_{u' \in e(u)} w(u'), \quad (11.1)$$

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

11.3.3 *S-Measure*

S-measure (Sakai et al. 2011a) was the primary evaluation metric at NTCIR-9 1CLICK-1 and NTCIR-10 1CLICK-2. Letting M be a set of iUnits identified in a summary, *S*-measure is defined as

$$S\text{-measure} = \frac{1}{\mathcal{N}} \sum_{u \in M} w(u) \max(0, 1 - \text{offset}(u)/L), \quad (11.2)$$

where $\mathcal{N}$ is a normalization factor, $w(u)$ is the weight of an iUnit u , L is a *patience* parameter, and $\text{offset}(u)$ is the offset of an iUnit u in the summary (more precisely, it is the number of characters between the beginning of the summary and the end of the iUnit). This measure basically represents the sum of the weight ($w(u)$) with offset-based decay ($1 - \text{offset}(u)/L$) for iUnits in a summary. Figure 11.2 illustrates *S*-measure computation with a simple example. As shown in the figure, the decay is assumed to decrease linearly with respect to the offset of an iUnit, and totally cancels the value of an iUnit appearing after L characters (the maximum function simply prevents the decay from being negative). Thus, the patience parameter can be interpreted as how many characters can be read by the user, or, alternatively, how much time the user can spend to read the summary when it is divided by the reading speed. For example, $L = 500$ in Fig. 11.2. If the reading speed is 500 characters per minute for average Japanese users, this patience parameter indicates that the user spends only a minute and leaves right after a minute passes. This corresponds to the fact that the decay factor becomes zero (or no value) after 500 characters.

The normalization factor $\mathcal{N}$ sets the upper bound so that *S* ranges from 0 to 1, and is defined as

$$\mathcal{N} = \sum_{u \in U} w(u) \max(0, 1 - \text{offset}^*(v(u))/L), \quad (11.3)$$

where U is a set of all iUnits and $\text{offset}^*(v(u))$ is the offset of the vital string of an iUnit u in *Pseudo Minimal Output* (PMO), which is an ideal summary artificially created for estimating the upper bound. The PMO was obtained by sorting all vital strings by $w(u)$ (first key) and $|v(u)|$ (second key) and concatenating them. Note that this procedure of generating an ideal summary may not be optimal, yet it is not a serious problem in practice as discussed in the original paper (Sakai et al. 2011a).

Finally, the original notation of *S*-measure is shown below, though it is obviously equivalent to Eq. 11.2:

$$S\text{-measure} = \frac{\sum_{u \in M} w(u) \max(0, L - \text{offset}(u))}{\sum_{u \in U} w(u) \max(0, L - \text{offset}^*(v(u)))}, \quad (11.4)$$

Fig. 11.2 Illustration of S-measure computation. The x-axis represents the number of characters read by the user, and y-axis represents the offset-based decay $(\max(0, 1 - \text{offset}(u))/L)$ with $L = 500$. The x-axis can also be interpreted as reading time indicated in the parentheses when the reading speed is 500 characters per minute. The textual output located at the bottom includes three iUnits u_1 , u_2 , and u_3 . The position of iUnits is aligned to the x-axis and their offsets are 125, 250, and 500, respectively. Their weight is 1 for simplicity. S-measure for this textual output can be computed as $S\text{-measure} = \frac{1}{N} (1 \cdot 0.75 + 1 \cdot 0.50 + 1 \cdot 0.00) = \frac{1}{N} \cdot 1.25$

11.3.4 M-Measure

M-measure (Kato et al. 2016a) was the primary evaluation metric at NTCIR-11 MobileClick-1 and NTCIR-12 MobileClick-2, which was proposed for two-layered summaries.

Intuitively, a two-layered summary is good if: (1) The summary does not include non-relevant iUnits in the first layer; (2) The first layer includes iUnits relevant for all the intents; and (3) iUnits in the second layer are relevant for the intent that links to them.

To be more specific, the following choices and assumptions were made for evaluating two-layered summaries:

- Users are interested in one of the intents $i \in I_q$ by following the intent probability $P(i|q)$, where I_q is a set of intents for query q .
- Each user reads a summary following these rules:
 1. The user starts to read a summary from the beginning of the first layer.
 2. When reaching the end of a link l_i which interests a user with intent i , the user clicks on the link and starts to read its second layer s_i .
 3. When reaching the end of the second layer s_i , the user goes back to the end of the link l_i and continues reading.
 4. The user stops after reading no more than L characters.

- The weight of iUnits is judged per intent. Therefore, an iUnit is important for a user but may not be important for another user.
- The utility of text read by a user is measured by U-measure proposed by Sakai and Dou (2013), which consists of a position-based gain and a position-based decay function.
- The evaluation metric for two-layered summaries, *M*-measure, is the expected utility of text read by different users.

These choices and assumptions derive all possible *trailtexts* and their probability in a two-layered summary. A trailtext is a concatenation of all the texts read by a user, and can be defined as a list of iUnits and links consumed by the user. According to the user model described above, a trailtext of a user who is interested in intent i can be obtained by inserting a list of iUnits in the second layer s_i after the link of l_i . More specifically, given the first layer $\mathbf{f} = (u_1, \dots, u_{j-1}, l_i, u_j, \dots)$ and second layer $\mathbf{s}_i = (u_{i,1}, \dots, u_{i,|s_i|})$, trailtext $\mathbf{t}_i$ of intent i is defined as follows: $\mathbf{t}_i = (u_1, \dots, u_{j-1}, l_i, u_{i,1}, \dots, u_{i,|s_i|}, u_j, \dots)$. An example of trailtexts in a two-layered summary is shown in Fig. 11.3.

M-measure, an evaluation metric for the two-layered summarization, is the expected utility of text read by users:

$$M = \sum_{\mathbf{t} \in T} P(\mathbf{t})U(\mathbf{t}), \quad (11.5)$$

Fig. 11.3 Example of trailtexts in a two-layered summary. Suppose links l_1 and l_2 are interesting for users with intents 1 and 2, respectively. All the users start to read the summary from the beginning of the first layer and read iUnits u_1 and u_2 . A user with intent 1 clicks on link l_1 , reads the iUnits in the second layer $\mathbf{s}_1$, and goes back to the first layer for reading the rest. A user with intent 2 does not click on link l_1 but clicks on link l_2 , reads the iUnits in $\mathbf{s}_2$, and returns to the first layer. These different trails result in different trailtexts shown at the bottom of the figure

where T is a set of all possible trailtexts, $P(\mathbf{t})$ is a probability of going through a trailtext $\mathbf{t}$, and $U(\mathbf{t})$ is the U-measure score of a trailtext $\mathbf{t}$.

For simplicity, a one-to-one relationship between links and intents was assumed in NTCIR-12 MobileClick-2. Therefore, there is only a relevant link and a trailtext for each intent. It follows that the probability of each trailtext being generated is equivalent to the probability of the corresponding intent, *i.e.*, $P(\mathbf{t}_i) = P(i|q)$ where $\mathbf{t}_i$ denotes a trailtext read by users with intent i . Then, M-measure can be rewritten as

$$M = \sum_{i \in I_q} P(i|q) U_i(\mathbf{t}_i). \quad (11.6)$$

where $U_i(\mathbf{t}_i)$ is the U-measure score of a trailtext $\mathbf{t}_i$ for users with intent i .

The computation of U-measure (Sakai and Dou 2013) is the same as that of S-measure except for the normalization factor and definition of the weight. U-measure is defined as follows:

$$U_i(\mathbf{t}) = \frac{1}{\mathcal{N}} \sum_{j=1}^{|\mathbf{t}|} g_i(u_j) d(u_j), \quad (11.7)$$

where $g_i(u_j)$ is the weight of iUnit u_j in terms of intent i , d is a position-based decay function, and $\mathcal{N}$ is a *constant* normalization factor ($\mathcal{N}=1$ in NTCIR MobileClick). Note that a link in the trailtext is regarded as a non-relevant iUnit for the sake of convenience. The position-based decay function is the same as that of S-measure:

$$d(u) = \max(0, 1 - \text{offset}(u)/L). \quad (11.8)$$

11.4 Outcomes of NTCIR 1CLICK and MobileClick

This section highlights the outcomes of NTCIR 1CLICK and MobileClick. We first present the results of each task and then discuss their potential impacts.

11.4.1 Results

Table 11.4 shows the number of participants and submissions at each NTCIR task. While the first round of 1CLICK and MobileClick failed to attract many participants, the second round of each received a sufficient number of submissions from ten or more teams. Due to a small number of participants, we only summarize results from 1CLICK-2 and MobileClick-2.

Table 11.4 The number of participants and submissions at each NTCIR task

Year	NTCIR	Task	# of participants	# of submissions
2010	9	1CLICK-1	3	10
2011	10	1CLICK-2	10	38 (for the Main task)
2013	11	MobileClick-1	4	24 (for retrieval) & 11 (for summarization)
2014	12	MobileClick-2	12	37 (for retrieval) & 29 (for summarization)

The NTCIR-10 1CLICK-2 results showed that simple use of search engine snippets and the first paragraph of Wikipedia articles outperformed more sophisticated approaches for both of the English and Japanese queries. Those simple approaches were particularly effective for celebrity query types, while they were not for the other types such as local queries (Kato et al. 2013b).

The NTCIR-12 MobileClick-2 task results showed that some participants' runs outperformed the baselines. Since the MobileClick task required systems to group iUnits relevant to the same intent, some teams proposed effective methods to measure the similarity between intents and iUnits, and achieved significantly better results than baselines. For example, one of the top performers used word embedding for measuring the intent-iUnit similarity, and another team proposed an extension of topic-sensitive PageRank for the summarization task. Per-query analysis at MobileClick-2 also suggested that celebrity query types were easy, while local and Q&A types of queries are difficult for both baselines and participants' systems (Kato et al. 2016b).

11.4.2 Impacts

An evaluation metric for summaries, ranked lists, and sessions, U-measure, was proposed by Sakai and Dou (2013). As they explained, U-measure was inspired by S-measure and is a generalization of S-measure. U-measure was further extended to the evaluation of customer-helpdesk dialogues by Zeng et al. (2017).

Luo et al. (2017) proposed *height-biased gain* (HBG), an evaluation metric for mobile search engine result pages. HBG is computed by summing up the product of weight and decay that are both modeled in terms of result height in mobile search engine result pages. As the authors mentioned in their paper, U-measure is one of the evaluation metrics that inspired HBG.

Arora and Jones (2017a,b) adapted the definition of iUnits for their study on identifying useful and important information and how people perceive information.

In commercial search engines, *direct answers* or *featured snippets* have become an important part of the search engine result page. This functionality presents a text that answers a question given as a query, just like the textual output of NTCIR 1CLICK. As of May 2019, it seems that they only show a part of a webpage and do

not summarize multiple webpages. The evaluation methodology of NTCIR 1CLICK and MobileClick could be potentially useful when direct answers are composed from multiple webpages and need to be evaluated in detail.

11.5 Summary

This chapter introduced the earliest attempts toward test-collection-based evaluation for information access with smartphones, namely, NTCIR 1CLICK and MobileClick. Those campaigns aimed to develop an IR system that outputs a single, short text summary for a given query, which is expected to fit a small screen and to satisfy users' information needs without requiring much interaction. This chapter mainly discussed the novelty of the evaluation methodology used in those evaluation campaigns by contrasting it with ordinary summarization evaluation. Moreover, the potential impacts of NTCIR 1CLICK and MobileClick were discussed as well.

Acknowledgements I really thank Tetsuya Sakai for his long-term contributions to the NTCIR 1CLICK and MobileClick tasks. A large fraction of the technologies introduced in this chapter was invented by or inspired by him. I also would like to thank the other task organizers of NTCIR 1CLICK and MobileClick, Young-In Song, Virgil Pavlu, Matthew Ekstrand-Abueg, Takehiro Yamamoto, Mayu Iwata, Hajime Morita, and Sumio Fujita. I would like to express deep gratitude to all the participants in the NTCIR-9 1CLICK-1, NTCIR-10 1CLICK-2, NTCIR-11 MobileClick-1, and NTCIR-12 MobileClick-2 tasks. Finally, I would like to express my special thanks to Stefano Mizzaro, who reviewed the initial version of this chapter and provided many useful suggestions.

References

- Arora P, Jones GJ (2017a) How do users perceive information: Analyzing user feedback while annotating textual units. In: Proc. of the Second Workshop on Supporting Complex Search Tasks, pp 7–11
- Arora P, Jones GJ (2017b) Identifying useful and important information within retrieved documents. In: CHIIR, pp 365–368
- Babko-Malaya O (2008) Annotation of nuggets and relevance in gale distillation evaluation. In: LREC
- Carbonell JG, Goldstein J (1998) The use of MMR and diversity-based reranking for reordering documents and producing summaries. In: SIGIR, pp 335–336
- Crestani F, Mizzaro S, Scagnetto I (2017) Mobile information retrieval. Springer, Berlin
- Dang HT, Kelly D, Lin JJ (2007) Overview of the TREC 2007 question answering track. In: TREC
- Hovy EH, Lin CY, Zhou L, Fukumoto J (2006) Automated summarization evaluation with basic elements. LREC 6:604–611
- Kato MP, Ekstrand-Abueg M, Pavlu V, Sakai T, Yamamoto T, Iwata M (2013a) Overview of the NTCIR-10 1CLICK-2 Task. In: NTCIR-10 conference, pp 243–249
- Kato MP, Sakai T, Yamamoto T, Iwata M (2013b) Report from the NTCIR-10 1click-2 Japanese subtask: baselines, upperbounds and evaluation robustness. In: SIGIR, pp 753–756
- Kato MP, Ekstrand-Abueg M, Pavlu V, Sakai T, Yamamoto T, Iwata M (2014) Overview of the NTCIR-11 mobileClick task. In: NTCIR-11 Conference

- Kato MP, Pavlu V, Sakai T, Yamamoto T, Morita H (2016a) Two-layered summaries for mobile search: does the evaluation measure reflect user preferences? In: *Proceeding of the seventh international workshop on evaluating information access (EVIA 2016)*, pp 29–32
- Kato MP, Sakai T, Yamamoto T, Pavlu V, Morita H, Fujita S (2016b) Overview of the NTCIR-12 mobileClick-2 task. In: *NTCIR-12 conference*
- Li J, Huffman S, Tokuda A (2009) Good abandonment in mobile and PC internet search. In: *SIGIR*, pp 43–50
- Lin CY (2004) ROUGE: A package for automatic evaluation of summaries. In: *Text summarization branches out: proceedings of the ACL-04 workshop*, pp 74–81
- Luo C, Liu Y, Sakai T, Zhang F, Zhang M, Ma S (2017) Evaluating mobile search with height-biased gain. In: *SIGIR*, pp 435–444
- Nenkova A, McKeown K (2011) Automatic summarization. *Found Trends® Inf Retrieval* 5(2–3):103–233
- Nenkova A, Passonneau R, McKeown K (2007) The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Trans Speech Lang Process (TSLP)* 4(2):4
- Sakai T, Dou Z (2013) Summaries, ranked retrieval and sessions: a unified framework for information access evaluation. In: *SIGIR*, pp 473–482
- Sakai T, Kato MP (2012) One click one revisited: enhancing evaluation based on information units. In: *AIRS*, pp 39–51
- Sakai T, Kato MP, Song YI (2011a) Click the search button and be happy: evaluating direct and immediate information access. In: *CIKM*, pp 621–630
- Sakai T, Kato MP, Song YI (2011b) Overview of NTCIR-9 ICLICK. In: *NTCIR-9*, pp 180–201
- Tombros A, Sanderson M (1998) Advantages of query biased summaries in information retrieval. In: *SIGIR*, pp 2–10
- Voorhees EM (2003) Overview of the TREC 2003 question answering track. *TREC 2003*:54–68
- Zeng Z, Luo C, Shang L, Li H, Sakai T (2017) Test collections and measures for evaluating customer-helpdesk dialogues. In: *Proceedings of the eighth international workshop on evaluating information access (EVIA 2017)*, pp 1–9

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.