#####

Dear NTCIR-10 PC,

Thank you very much for your questions on 1CLICK-2.

Please find our answers below.

(1) Are you familiar with the nugget-based evaluation work that was done in the DARPA GALE program (e.g., http://www.springerlink.com/content/vw1q605402757744/)? If so, what can be learned from that work, both about evaluation and about system design, that might help to accelerate progress in this track.

Yes we are aware of the general framework of the DARPA GALE nugget evaluation.

At 1CLICK-1, we only had a brief nugget creation policy document:

http://research.microsoft.com/en-us/people/tesakai/nuggetpolicy.pdf

but as we mentioned in our 1CLICK-2 proposal,

we will devise a more systematic way of creating nuggets.

We believe that previous work such as this will be useful for it:

http://www.lrec-conf.org/proceedings/lrec2008/pdf/909\_paper.pdf

In particular, we can revise our nugget creation procedure based on their methodology, for fixing the granularity of nuggets, which is one of the main problems in 1CLICK@NTCIR-9.

On the other hand, we stress that we are not repeating what DARPA GALE did.

In the DARPA GALE nugget evaluation, nuggets were created from the submitted runs

in a bottom-up manner, and nugs (nugget clusters) were obtained from them.

Then nug-based recall and precision were computed.

Thus, just like other nugget-based evaluations (in summarisation and in QA),

the evaluation method is position-agnostic.

In contrast, 1CLICK defines important nuggets prior to collecting runs

(although we do allow for later revisions), and

requires systems to present important pieces of information first

and to minimise the amount of text the user has to read.

The user (e.g. in a mobile environment) has (say) only one minute to gather information.

So any given nugget becomes worthless after one minute.

As Figure 16 (page 18) of the 1CLICK-1 overview paper shows,

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-1CLICK-SakaiT.pdf

we can penalise systems that present the relevant information

at the very bottom of the output window.

Previous nugget-based evaluations, which disregarded the nugget position information,

cannot spot such practical problems.

Thus, while we can learn a lot from previous nugget-based evaluations,

we are confident that we will learn something new and useful.

Please note also that Virgil Pavlu has joined the organising team

to explore other new possibilities in nugget-based evaluation at 1CLICK-2.

(2) What is known about the reusability of annotations created in this track in the presence of Web link rot, the diversity of ways of expressing the same content that is possible, and prior work on automated evaluation of QA systems?

- web link rot

We are aware that the 1CLICK test collection is not entirely reusable,

and that at least some revisions are required for non-participants to use the data set.

We have discussed this a little in page 10 of the following paper:

http://research.microsoft.com/en-us/people/tesakai/cikmfp0040-sakai.pdf

While our weakness in terms of reliability arises mainly from the use of the live web,

we believe that this also has some advantages, because the live web is the reality.

Exploring methods for finding information from a small frozen corpus

may be worthless if there are much more useful sources out on the web.

As our approach of 1CLICK-1 had pros and cons,

for 1CLICK-2, we plan to release a small, frozen set of web data

for participants to share, in order to encourage more repeatable and

component-based evaluation.

- diversity of expressing the same content

In contrast to the DARPA GALE approach

where multiple nuggets form an equivalence class called nugs,

the 1CLICK approach does not require a collection

of different expressions representing the same piece of information.

Instead, assessors manually compare the system output and the list

of gold-standard nuggets, and identify matching nuggets as well as their

positions. Thus it is the assessors who bridge the gap between

the system's expression and the expression in the gold standard set.

We plan to make the nugget match evaluation interface

publicly available, so that non-participants can utilise

the 1CLICK evaluation framework, possibly with their own data.

- prior work on automated evaluation of QA

1CLICK test collections may indeed be less reusable than

QA test collections that used a "closed" document collection

as the web is dynamic and so are search engine APIs.

However, as we have argued, this also gives us the advantage of

being able to tackle the real problems.

For example, while the QA test collections from past NTCIRs

are valuable(\*), their corpora are news articles.

The 1CLICK participants faced many practical problems

that previous QA participants did not,

such as extracting sentences from heterogeneous HTML pages

and handling conflicting information from

multiple different sources

(e.g. wikipedia vs official profile site).

(\*) One 1CLICK participant actually utilised the NTCIR ACLIA (QA)

test collections to design their summarizer.

(3) Are there opportunities for synergy with the RITE track? If so, how might such synergies best be achieved?

Indeed, we do believe that there is a connection,

and feel that a future collaborations would be possible.

1CLICK evaluation requires manual nugget match evaluation,

but if textual entailment technology matures, this

could partially replace the matching process.

Section 2.2 (page 2) of the following paper discusses this a little:

http://research.microsoft.com/en-us/people/tesakai/cikmfp0040-sakai.pdf

We do feel, on the other hand,

that it would be too early for RITE and 1CLICK to try to collaborate heavily for NTCIR-10:

both tasks are still in the exploration phase.

(4) Distributed annotation for evaluation by participants would seem to raise some concerns about quality control. Are such concerns justified? Why or why not? If they are, even if only to a modest degree, how will effects be measured and how will risks be mitigated?

We exchanged system outputs among participants and used a randomised evaluation environment,

so we are not too concerned about this issue.

At 1CLICK-1, we used two assessors per system output, and

we had 100 queries in total.

We compared the system ranking results

based on one assessor,

and on the intersection and the union of

the two assessors' identified nuggets.

As page 11 of the 1CLICK overview paper shows,

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-1CLICK-SakaiT.pdf

the evaluation outcomes are similar.

Moreover, the Cohen’s Kappa values between two different assessors were generally high: they ranged from 0.68 to 0.88.

FYI, page 16 of the same paper

shows some close examinations of

inter-assessor disagreements:

there ARE some misses and false alarms,

but their impact on the overall results is limited.

(5) Multi-document summarization competes with the human ability to rapidly skim documents, which (with diversity ranking) can be both rather efficient and very effective at conveying context (which is a serious challenge for extractive summarization). The same is not true, however, for speech, where skimming is impractical. Have you considered working together with the SpokenDoc track to tailor a task for their content? Why or why not?

We have not considered any collaboration with SpokenDoc as our interests lie elsewhere.

To be more specific, we are interested mostly in mobile search situations

where the output screen size is small and the time that can be spent for

gathering information is limited. We want to build systems that

gather important pieces of information from multiple textual sources and

present them concisely. We do not believe that the task of

retrieving spoken lectures fits well in our scenario.

On the other hand, speech \*input\* would be a realistic setting for 1CLICK,

as speech input mobile search is quickly becoming a reality. We are also interested in

a form of "GeoTime" mobile search, where the query is not an explicit string but a context

such as longitude and latitude or some sensor data.

In fact we will include a LOCAL type of query in NTCIR-10, which comprises a location name and a type of facility (e.g. “tokyo hotel”), and is somewhat connected to the GeoTime task.

However we do feel that it is too early to tackle these problems within the 1CLICK framework.

(6) The motivation for this task seems to overlap to some degree with the motivation for the Diversity Ranking task of the Intent track. Are there potential synergies with that task? If so, how would those synergies best be achieved? If not, what differences prevent effective coordination or collaboration between these tasks?

Yes, there is a lot of overlap among the INTENT and 1CLICK organisers and we do have some common interests.

We see this as an advantage - it may be possible to borrow some ideas for diversity evaluation from INTENT

and apply it to 1CLICK, or conversely, if we successfully establish a nugget-based IR evaluation framework

that applies not only to 1CLICK but also to general IR, this could be applied to INTENT evaluation.

We may also share some topics across these tasks.

(7) It has been observed that extractive multi-document summarization bears such great similarity to list QA that the two are essentially indistinguishable. What prior work has been done on evaluation of list QA and on multi-document summarization in venues such as TREC, CLEF, TAC, NTCIR, and GALE, and in the context of that answer what further advances can reasonably be expected from now continuing 1-Click?

Please see our answers to Question (1).

In addition, please note the following two points:

1. Some of the previous nugget-based evaluations rely on automatic matching between the system output and the gold standard

(based on word/character overlaps etc). We feel that automatic matching alone is insufficient

for evaluating concise output where abstractive summarization will play an important role.

2. Some of the previous nugget-based evaluations rely on an arbitrary "allowance" to define nugget precision.

This in effect means that every nugget requires a fixed amount of space, which is probably not a good assumption.

While our 1CLICK evaluation framework is far from perfect, we believe that it is a useful departure from

existing work: we try to approximate the minimum amount of space required for each nugget.

(8) In NTCIR-9 this track had fewer participants than the NTCIR-10 track proposal has organizers. This raises cost-benefit concerns. What degree of confidence do you have that a substantial number of participants can be attracted in NTCIR-10, and what is your basis for that confidence?

At 1CLICK-1, we had three organisers and other twelve participants, as follows:

1CLICK-1 Organisers: Sakai, Kato, Song (3)

Participants, excluding the organisers:

Zhao, Tsukuda, Shoji, Yamamoto, Ohshima, Tanaka

Morita, Makino, Takamura, Okumura

Naoki Orii (12)

Young-In Song, the Korean organiser, will run the Korean subtask

and he has already found a few potential 1CLICK-2 participants from Korea.

Virgil Pavlu from Northeastern University will run the English subtask

and is likely to bring in new participants from the US at least.

As for Japan, we will work harder in marketing. As 1CLICK has been promoted from

a pilot task to a core task, and has demonstrated the feasiblity and usefulness

of the new evaluation framework, we are confident that 1CLICK-2 will attract more participants.

Synergy between INTENT and 1CLICK:

we will encourage INTENT-2 participants to also participate in 1CLICK, and vice versa.

(9) Have you considered a multilingual task in which answers would need to be deduplicated across different languages? Why or why not? If this is something that might be done in NTCIR-10 or some subsequent evaluation, have you considered the possibility that you might coordinate in some way with the evaluation of that task that is planned for the DARPA BOLT program in the USA?

Crosslingual 1CLICK would be interesting and

we are already discussing whether we can share some topics

across Japanese, Korean and English.

However,

we feel we know too little about monolingual 1CLICK

and that introducing a crosslingual subtask is too early.

Our current focus is on handling real mobile queries,

and real mobile queries are quite different across countries

(e.g. see J. Li, S. Huffman, and A. Tokuda. Good abandonment in

mobile and PC internet search. In Proceedings of ACM

SIGIR 2009).

So it might be better to handle the crosslingual and multilingual issues outside 1CLICK,

possibly at NTCIR-11 as a new task.

(10) What Is known about the reusability of nugget-based evaluation from prior work in GALE and list QA? Is there a solid basis for believing that the investment in test collection development can be amortized over future uses of the collection? If not, is there a means for gaining confidence that that is the case? If not, is there a reason to run this task despite not having such an assurance?

In general, we believe that nuggets constructed based on a limited set of submitted runs

are not reusable. Compared to the case with document-relevance-based IR test collections,

the incompleteness issue is too serious.

The 1CLICK nugget construction scheme is more top-down and does not directly depend

on the submitted runs. It aims at collecting time-insensitive truths.

(But many "truths" do change over time, so automatic truth maintenance for nuggets

may be a useful research problem.)On the other hand, our use of the live web and search engine APIs does hurt repeatability.

We have already discussed above our plan to share some intermediate data across participants

to improve the situation.

Even if the 1CLICK data sets are not reusable, however,

we believe that the task has many merits, as discussed earlier.

In addition, please note that our legacy such as the nugget evaluation interface

will be reusable, possibly with some updates.

#####