Vol.2015-DBS-161 No.5
Vol.2015-IFAT-119 No.5
2015/8/5

Topic Set Size Design with the Evaluation Measures for Short Text Conversation

Tetsuya Sakai[1,a)]

Abstract: Short Text Conversation (STC) is a new NTCIR task which tackles the following research question: given a microblog repository and a new post to that microblog, can systems reuse an old comment from the respository to satisfy the author of the new post? The official evaluation measures of STC are *normalised gain at 1* (nG@1), *normalised expected reciprocal rank at 10* (nERR@10), and P^+, all of which can be regarded as evaluation measures for navigational intents. In this study, we apply the *topic set size design* technique of Sakai to decide on the number of test topics, using variance estimates of the above evaluation measures. Our main conclusion is to create 100 test topics, but what distinguishes our work from other tasks with similar topic set sizes is that we know what this topic set size means from a statistical viewpoint for each of our evaluation measures. We also demonstrate that, under the same set of statistical requirements, the topic set sizes required by nERR@10 and P^+ are more or less the same, while nG@1 requires more than twice as many topics. To our knowledge, our task is the first among all efforts at TREC-like evaluation conferences to actually create a new test collection by using this principled approach.

Keywords: evaluation, measures, microblog, power, statistical significance, test collections.

1. Introduction

Short Text Conversation (STC)[*1] is a new

NTCIR task which tackles the following research question: given a microblog repository and a new post to that microblog, can systems *reuse* an old comment from the respository to satisfy the author of the new post? For each new post, systems are expected to output a ranked list of past comments that are *coherent* with respect to the original post and *useful* from the viewpoint of the author of the post. For example, given a post "The first day in Hawaii. Watching the sunset at the balcony with a big glass of wine in hand," comments such as "Enjoy it & don't forget to share your photos!" and "How long are you going to stay there?" are coherent, and could also be considered useful to the author in Hawaii[*2]. We view this as a first small step towards developing a system that can interact effectively with the user in natural language; the objective of STC is to quantify how far we can go using a purely IR-oriented approach that does not involve natural language generation. While retrieving and ranking coherent and useful comments is different from the traditional IR task of ranking items that are *relevant* to an information need, we expect that various wisdoms of IR such as the pooling technique and graded relevance measures will be applicable to, and highly useful for, this task.

In the first round of STC at NTCIR-12[*3], a Chinese *Weibo*[*4] corpus will be used. Weibo currently has over 40 million users, and is very much like Twitter[*5] in terms of user experience: just like Twitter, each Weibo "tweet" has the length limit of 140 characters, although 140 characters in Chinese can be significantly more informative than 140 characters in English, as the Chinese characters are ideograms with no spaces between words[*6]. Table 1 shows the structure of the STC test collection: (a) the repository of "old" posts and their comments; (b) labelled post-comment pairs for training; and (c) test data that will be contructed as an outcome of the STC task. Note that the posts in our training and test data were sampled from outside the repository to be treated as "new" posts, while the comments in these data sets are from the repository, which are regarded as "reused" comments. That is to say, for every labelled post-comment pair in the STC test collection, the comment was originally a response to some other post.

Table 1 STC test collection.

(a) Repository	#posts	196,395
	#post-comment pairs	5,648,128
(b) Training data	#posts	225
	#post-comment pairs (labelled)	6,017
(c) Test data	#posts	TBD
	#post-comment pairs (labelled)	TBD

The training data labels were obtained as described in the aforementioned `arxiv` paper. Briefly, for each of our training post, we searched the repository using three simple algorithms, and pooled the top 10 comments from each run. The comments

[1] Waseda University, Japan
[a)] tetsuyasakai@acm.org

[*1] `http://ntcir12.noahlab.com.hk/stc.htm`
[*2] Examples taken from the `arxiv` paper by Ji, Lu and Li: `http://arxiv.org/pdf/1408.6988.pdf`.
[*3] `http://research.nii.ac.jp/ntcir/ntcir-12/`
[*4] `http://weibo.com`
[*5] `http://twitter.com`
[*6] The minimum/average/maximum lengths of the 196,395 posts in the repository are 10/32.5/140, respectively. Whereas, after translating them into English using machine translation, the corresponding lengths are 11/115.7/724. This suggests that a Chinese tweet can be 3-5 times as informative as an English one.

in the depth-10 pools were then manually assessed from multiple viewpoints to form graded "relevance" data, with relevance grades L0 (not relevant), L1 (relevant) and L2 (highly relevant)[*7]. In the present study, we evaluate six runs based on the training data labels in order to estimate the within-system variances of several evaluation measures and thereby determine the number of test topics (i.e., posts) in a principled way. While our training data labels are probably highly incomplete and biased, note that we are running the STC task exactly because we want to create a reliable STC test collection with a test topic set with post-comment labels obtained via a pooling of a variety of runs. See Section 5 for more discussions.

The official evaluation measures of STC are *normalised gain at 1* (nG@1) [15][*8], *normalised expected reciprocal rank at 10* (nERR@10) [2], and P^+ [11], all of which can be regarded as evaluation measures for *navigational* intents[1]. In this study, we apply the *topic set size design* technique of Sakai [13], [14] to decide on the number of test topics, using variance estimates of the above evaluation measures. Our main conclusion is to create 100 test topics, but what distinguishes our work from other tasks with similar topic set sizes is that we know what this topic set size means from a statistical viewpoint for each of our evaluation measures. We also demonstrate that, under the same set of statistical requirements, the topic set sizes required by nERR@10 and P^+ are more or less the same, while nG@1 requires more than twice as many topics. To our knowledge, our task is the first among all efforts at TREC-like evaluation conferences to actually create a new test collection by using this principled approach.

2. Related Work

2.1 Evaluation Tasks Related to STC

As the STC task requires participating systems to produce a ranked list of comments given a Weibo post, it is very similar to traditional TREC ad hoc tracks [19], in terms of input/output specifications and the test collection construction procedure. A post is like a TREC topic, and comments are like target documents; instead of retrieving relevant documents, STC systems are expected to retrieve coherent and useful comments. Just like TREC, the STC runs will be pooled, with a pool depth of 10, and graded "relevance" assessments will be conducted using multiple assessors for judging each comment.

In terms of document type, STC resembles the TREC Microblog track which uses Twitter data. At the TREC 2011 and 2012 Microblog tracks, a collection comprising 16 million tweets were used, but only tweet IDs were distributed to participating teams and each team had to download the actual data for themselves. This meant that the different downloads were not strictly identical. Whereas, from the TREC 2013 Microblog track, "Evaluation as a Service" was introduced to handle over 243 million tweets via search APIs [7], which meant that participating teams did not have direct access to the actual data. In contrast, while the STC Weibo collection is relatively small (See Table 1), the entire data set is distributed to each participating team for research purposes, in a way similar to the "TREC disks" [19].

In terms of task, STC is related to question answering (QA) tasks such as the TREC QA track [19], the NTCIR ACLIA (Advanced Crosslingual Information Access) task [8], and the NTCIR QALab task [18]. In particular, the NTCIR CQA (Community QA) task [15] is related to STC in terms of both document type and task: CQA used the Yahoo! Chiebukuro (Japanese Yahoo! Answers) data, and the task was to find the answer to a question that was selected by the questioner as the "best answer." The most important distinction between these QA-related tasks and STC is that an STC post is not necessarily a question, and therefore that each comment to the post is not necessarily an answer. For example, in the example given in Section 1, note that one of the *comments* is a question: "How long are you going to stay there?"[*9].

2.2 Problems and Approaches Related to STC

Research on modelling human-computer dialogues started over half a century ago [21], but the recent advent of social media such as Twitter has revitalised this area using new approaches. STC is the simplest form of human-computer dialogues that deals with one post-comment pair at a time, and statistical modelling of STC and related tasks based on large scale social media corpora has become possible. For example, Ritter, Cherry and Dolan [10] utilised the Twitter data to study the feasiblity of *generating* a comment to a given post, by regarding the transformation from a post to a comment as a statistical translation problem. This is in contrast to the STC problem setting where systems are expected to *reuse* comments from a social media repository. Using Twitter and live-journal data, Jafarpour and Burges [5] tackled a problem they refer to as *learning to chat*, which is very similar to STC in that past comments are retrieved for reuse, although they mention in their paper that the retrieved comment should then be altered prior to presentation to the author of the new post. They propose a three-stage approach to ranking past comments, and also a mechanism for collecting high-quality training data from users. Higashinaka *et al.* [4] learn a conversational model from post-comment pairs (or "Two-Tweet exchanges"), and report that the learned model is comparable in effectiveness to one that utilises longer exchanges as training data.

We are hoping that many research groups that are tackling related problems such as the ones mentioned above will participate in the NTCIR-12 STC task. We shall report on the outcome of STC in our NTCIR-12 overview paper in 2016, where we hope to clarify what kind of techniques are effective for this relatively simple form of human-computer dialogue.

2.3 Topic Set Size Design

Sakai [13], [14] showed three statistically motivated methods for determining the topic set size for a test collection to be built: one based on the paired *t*-test, one based on one-way ANOVA

[*7] While the present study uses the post-comment labels collected as described in the `arxiv` paper, we have since then revised the labelling criteria in order to clarify several different axes for labelling, including coherence and usefulness. The new labelling scheme will be used to revise the training data labels as well as to construct the official test data labels.

[*8] nG@1 is sometimes referred to as nDCG@1; however, note that neither discounting ("D") nor cumulating gains ("C") is applied at rank 1.

[*9] Given an input remark "Men are all alike," ELIZA, the rule-based system developed in the 1960s, could respond: "IN WHAT WAY?" [21]

and one based on confidence intervals (CIs). In the present study, we use Sakai's ANOVA-based Excel tool[*10] as this method can consider comparison of $m(\geq 2)$ systems and is the most general. Sakai demonstrated that the ANOVA-based method with $m = 2$ and the *t*-test-based method give similar results, and also that the ANOVA-based method with $m = 10$ can be used instead of the CI-based method (See Section4.2).

Sakai's ANOVA-based tool requires the following input parameters to determine the required topic set size:

- α The probability of *Type I error* (detecting a difference that does not exist).
- β The probability of *Type II error* (missing a difference that actually exists).
- m**:** The number of systems that will be compared in one-way ANOVA ($m \geq 2$).
- *minD***:** The *minimum detectable range* [13], [14]. That is, whenever the performance difference between the best and the worst systems is *minD* or higher, we want to ensure a *statistical power* of $(1 - \beta)$ (i.e., the probability of detecting a difference that actually exists) given the significance level α.
- $\hat{\sigma}^2$**:** The estimated variance of a system's performance, under the *homoscedasticity* (i.e., equal variance) assumption [9], [13], [14]. That is, it is assumed that the scores of the i-th system obey $N(\mu_i, \sigma^2)$, where μ_i's differ while σ^2 is common to all systems. This variance known to be heavily dependent on the evaluation measure.

Sakai [13], [14] also describes simple ways to obtain $\hat{\sigma}^2$ for a particular evaluation measure, given a $n \times m$ topic-by-system matrix of scores x_{ij}, for system i and topic j. We use his variance estimation method based on one-way ANOVA: let the sample mean for system i be $\bar{x}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$; the population within-system variance can be estimated as:

$$\hat{\sigma}^2 = V_E = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\bullet})^2}{m(n-1)} . \quad (1)$$

3. Evaluation Measures for Short Text Conversation

The official evaluation measures of the STC task are graded-relevance IR evaluation measures for *navigational* intents [1]. This is because a human-computer conversation system that can respond naturally to a natural language post would usually require exactly one good comment. Below, we define the official measures and clarify the relationships among them. We compute these evaluation measures using the `NTCIREVAL` tool[*11].

3.1 nG@1

Let $g(r)$ denote the *gain* of a document (i.e., a comment) retrieved at rank r: throughout this paper, we let $g(r) = 2^2 - 1 = 3$ if the document is L2-relevant; $g(r) = 2^1 - 1 = 1$ if it is L1-relevant; $g(r) = 0$ if it is not relevant (i.e., L0). For a given topic (i.e., a post), an *ideal ranked list* is constructed by listing up all L2-relevant documents followed by all L1-relevant ones. Let $g^*(r)$ denote the gain of a comment at rank r in the ideal list. Normalised Gain at Rank 1 is defined as follows:

$$nG@1 = \frac{g(1)}{g^*(1)} . \quad (2)$$

This is a crude measure, in that it only looks at the top ranked document, and that, in our setting, it only takes three values: 0, 1/3 or 1.

3.2 nERR@10

Expected Reciprocal Rank (ERR) [2] is a popular measure with a *diminishing return* property: once a relevant document is found in the list, the value of the next relevant document in the same list is guaranteed to go down. Hence, the measure is suitable for navigational intents where the user does not want redundant information. ERR assumes that the user scans a ranked list from top to bottom, and that the probability that the user is satisfied with the document at rank r is given by $p(r) = \frac{g(r)}{2^H}$, where H denotes the highest relevance level for a test collection (2 in our case). Hence, in our setting, $p(r) = 3/4$ if the document at rank r is L2-relevant; $p(r) = 1/4$ if it is L1-relevant; $p(r) = 0$ if it is not relevant. The probability that the user reaches as far as rank r and then stops scanning the list (due to satisfaction) is given by:

$$Pr_{ERR}(r) = p(r)\prod_{k=1}^{r-1}(1 - p(k)) , \quad (3)$$

and the *utility* of the ranked list to the user who stopped at r is computed as $1/r$ (i.e., only the final document is considered to be useful). Therefore, ERR is defined as:

$$ERR = \sum_{r} Pr_{ERR}(r)\frac{1}{r} . \quad (4)$$

ERR is known to be a member of the *Normalised Cumulative Utility* (NCU) family [16], which is defined in terms of a stopping probability distribution over ranks ($Pr_{ERR}(r)$ in this case) and the utility at a particular rank ($1/r$ in this case).

As ERR is not normalised, it may be normalised using the aforementioned ideal list. Let $p^*(r)$ denote the stopping probability at rank r in an ideal list, let $Pr^*_{ERR}(r)$ be defined in a way similar to Eq 3. Normalised ERR at a cutoff l is given by:

$$nERR@l = \frac{\sum_{r=1}^{l} Pr_{ERR}(r)(1/r)}{\sum_{r=1}^{l} Pr^*_{ERR}(r)(1/r)} . \quad (5)$$

The primary measure of STC is nERR@10. Note that, when $l = 1$ in Eq. 5,

$$nERR@1 = \frac{Pr_{ERR}(1)}{Pr^*_{ERR}(1)} = \frac{p(1)}{p^*(1)} = \frac{g(1)/2^H}{g^*(1)/2^H} = \frac{g(1)}{g^*(1)} = nG@1 . \quad (6)$$

That is, nG@1 can alternatively be referred to as nERR@1.

3.3 P+

P+, proposed at AIRS 2006 [11], is another evaluation measure designed for navigational intents. Like ERR, it is a member of the NCU family. Given a ranked list, let r_p be the rank of the document that has the highest relevance level in that particular list (which may or may not be H, the highest relevance level for the entire test collection) *and* is closest to the top of the list. For

[*10] `http://www.f.waseda.jp/tetsuya/CIKM2014/samplesizeANOVA.xlsx`
[*11] `http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html`

example, if the ranked list has L2-relevant documents at ranks 2 and 5, and an L1-relevant document at rank 1, then $r_p = 2$; if the ranked list does not contain any L2-relevant documents but has L1-relevant document at ranks 3 and 5, then $r_p = 3$. The basic assumption behind P$^+$ is that no user will ever go beyond r_p: the *preferred rank*.

P$^+$ assumes that the distribution of users who will stop scanning the ranked list at a particular rank is uniform over all relevant documents at or above r_p. For example, if there is an L1-relevant document at rank 1 and an L2-relevant document at rank $r_p = 2$, then it is assumed that 50% of users will stop at rank 1, and the other 50% will stop at rank 2. More generally, let $I(r) = 0$ if the document at rank r is not relevant and $I(r) = 1$ otherwise; the stopping probability at each relevant document at or above r_p is assumed to be[*12] $1/\sum_{r=1}^{r_p} I(r)$.

While ERR uses the *reciprocal rank* $(1/r)$ to measure the utility of a ranked list for users who stopped at rank r, P$^+$ employs the *blended ratio* $BR(r)$ just like *Q-measure* [16]:

$$BR(r) = \frac{\sum_{k=1}^{r} I(k) + \sum_{k=1}^{r} g(k)}{r + \sum_{k=1}^{r} g^*(k)} . \quad (7)$$

Note that *precision* based on binary relevance is given by $P(r) = \sum_{k=1}^{r} I(k)/r$, while *normalised cumulative gain* [6] based on graded relevance is given by $nCG(r) = \sum_{k=1}^{r} g(k)/\sum_{k=1}^{r} g^*(k)$. $BR(r)$ combines these two measures; the r in the denominator of Eq. 7 discounts documents based on ranks.

Finally, P$^+$ is defined as follows. If the ranked list does not contain any relevant documents, let $P^+ = 0$. Otherwise,

$$P^+ = \sum_r Pr_+(r)BR(r) = \frac{1}{\sum_{r=1}^{r_p} I(r)} \sum_{r=1}^{r_p} I(r)BR(r) . \quad (8)$$

Here, $Pr_+(r)$ denotes the aforementioned uniform stopping probability distribution over relevant documents ranked at or above rank r_p.

Consider a ranked list that contains one document only. If this document is not relevant, $P^+ = 0$ by definition. If it is relevant, then $r_p = 1$ and $I(1) = 1$, and therefore

$$P^+ = \frac{1}{I(1)} I(1)BR(1) = BR(1) = \frac{I(1) + g(1)}{1 + g^*(1)} = \frac{1 + g(1)}{1 + g^*(1)} , \quad (9)$$

which is very similar to the definition of nCG@1 (a.k.a. nERR@1). Also note that, regardless of the ranked list size, $P^+ = 1$ iff $r_p = 1$ *and* the top ranked document is one of the most relevant ones for that topic.

4. Experiments

This section reports on how we decided on the topic set size for the STC test topics (i.e., posts) using Sakai's ANOVA-based topic set size design tool [13], [14], the STC repository and the training data labels described in Table 1, and the aforementioned three official evaluation measures.

[*12] Note that *Average Precision* and *Q-measure* assume a uniform distribution over *all* relevant documents, so that the stopping probability each relevant document is $1/R$, where R is the total number of relevant documents [16].

4.1 Pilot Runs

As was mentioned in Section 2.3, topic set size design requires an estimate of the population within-system variance for a given evaluation measure. To obtain the variance estimate using Eq. 1, we created a topic-by-system matrix for each of the three evaluation measures using the $n = 225$ training topics from Table 1 and $m = 6$ pilot runs we created. Our pilot runs employ *learning-to-match* and *learning-to-rank* models as described in the aforementioned `arxiv` paper (See Section 1). Table 2 shows the combinations of features used to generate these runs, where the features used are:

Q2P Query-post similarity based on the vector space model. Here, "query" refers to the new post as an input to an STC system, whereas "post" refers to an old post in the repository. The basic assumption is that if these two posts are similar, then their comments will likely be exchangeable.

Q2C Query-comment similarity based on the vector space model. Again, "query" refers to the new post, while "comment" refers to one from the repository. The basic assumption is that a good comment contains words that are similar to those in the new post.

TransLM Translation-based language model for bridging the lexical gap between the query and candidate post-comment pairs, which Q2P and Q2C cannot handle. Word-to-word translation probabilities are estimated so that any word in a post or a comment can be translated with a non-zero probability into a semantically related query word.

TopicWord Topic word model for estimating the probability that each word in a post or comment is to do with the main topic or theme. Logistic regression with features such as term frequency, inverse document frequency, whether the word is a named entity, and whether the word occurs in the first (last) sentence is employed.

Table 2 also shows the mean performances of these runs for the training data, and Table 3 shows, for each run pair, the p-value obtained with the *randomised Tukey HSD test* for multiple comparison with $B = 5000$ trials using the `Discpower` tool[*13], as well as the *effect size* ES_{HSD} [12][*14]. However, these results should be regarded with a large grain of salt, because (a) the training data labels were contructed based on pooling only three runs and therefore may be highly incomplete and biased; and (b) the new six pilot runs have been tuned with these training data labels. The purpose of these runs in the present study is to estimate the within-system variances rather than performance comparisons. It can be observed, however, that introducing the TopicWord feature may actually hurt the mean performance (Compare Run2 and Run4), and that the effect of TransLM is not statistically significant (Compare Run2 and Run3, or Run4 and Run 5), even on the training data.

[*13] `http://research.nii.ac.jp/ntcir/tools/discpower-en.html`

[*14] The effect size here is essentially the difference between a system pair as measured in standard deviation units, after removing the between-system and between-topic effects.

Table 2 Six pilot runs used for obtaining $\hat{\sigma}^2$'s, and their mean performances on training data.

Run name	Features used	nERR@10	P+	nG@1
Run0	Q2P	.5839	.6050	.4015
Run1	Q2C	.6437	.6659	.4637
Run2	Q2P + Q2C	.6908	.7140	.5496
Run3	Q2P + Q2C + TransLM	.6913	.7149	.5318
Run4	Q2P + Q2C + TopicWord	.6866	.7095	.5392
Run5	Q2P + Q2C + TransLM + TopicWord	.6909	.7121	.5363

Table 3 p-values/effect sizes (ES_{HSD}) for pairwise comparisons of the six runs. p-values smaller than $\alpha = 0.05$ are shown in bold.

(a) nERR@10	Run1	Run2	Run3	Run4	Run5
Run0	**.004**/.3392	**.000**/.6065	**.000**/.6091	**.000**/.5829	**.000**/.6070
Run1	-	**.040**/.2673	**.037**/.2699	.076/.2438	**.040**/.2678
Run2	-	-	1.000/.0026	1.000/.0236	1.000/.0005
Run3	-	-	-	1.000/.0262	1.000/.0021
Run4	-	-	-	-	1.000/.0241
(b) P+	Run1	Run2	Run3	Run4	Run5
Run0	**.006**/.3450	**.000**/.6177	**.000**/.6231	**.000**/.5924	**.000**/.6073
Run1	-	.057/.2727	**.048**/.2781	.108/.2474	.075/.2622
Run2	-	-	1.000/.0054	1.000/.0253	1.000/.0104
Run3	-	-	-	.999/.0307	1.000/.0159
Run4	-	-	-	-	1.000/.0148
(c) nG@1	Run1	Run2	Run3	Run4	Run5
Run0	.194/.3528	**.000**/.8402	**.000**/.7393	**.000**/.7813	**.000**/.7645
Run1	-	**.014**/.4873	.106/.3865	.058/.4285	.071/.4117
Run2	-	-	.987/.1008	.998/.0588	.996/.0756
Run3	-	-	-	1.000/.0420	1.000/.0252
Run4	-	-	-	-	1.000/.0168

4.2 Topic Set Size Design Results

We created a 225×6 topic-by-system matrix for each of our evaluation measure based on `NTCIREVAL`, obtained the within-system variances using Eq. 1, and then used Sakai's ANOVA-based Excel tool with $(\alpha, \beta) = (0.05, 0.20)$, i.e., *Cohen's five-eighty convention* [3], which says that a Type I error is four times as serious as a Type II error. Table 4 shows the required topic set sizes given the minimum detectable range $minD = 0.05, \ldots, 0.20$ and the number of systems to be compared $m = 2, \ldots, 100$ for the three evaluation measures. It can be observed that the within-system variances of nERR@10 and P+ are very similar, and therefore that the required topic set sizes are also very similar under a given set of statistical requirements $(\alpha, \beta, minD, m)$. For example, if we are to compare $m = 10$ systems using one-way ANOVA and want to guarantee $(\alpha, \beta, minD) = (0.05, 0.20, 0.15)$, that is, if we want to guarantee 80% *statistical power* at 5% significance level whenever there is a difference of 0.15 or more between the best and the worst systems, P+ would require 89 topics, while nERR@10 would require 90 topics. Whereas, note that nG@1 would require as many as 211 topics under the same condition, due to the fact that it is a highly unstable measure.

Based on Table 4, we have decided to create a test set containing 100 posts for STC and release them to participating teams in November 2015. From the same table, the statistical implications of this decision under Cohen's five-eighty convention are as follows:

- If P+ or nERR@10 is used for evaluation, this test set would achieve a minimum detectable difference of 0.10 for comparing $m = 2$ systems[*15];
- If P+ or nERR@10 is used for evaluation, this test set would achieve a minimum detectable range of 0.15 for comparing $m = 10$ systems; also, this test set would be expected to make the confidence interval width of the difference between any systems be 0.15 or smaller [13], [14];
- If P+ or nERR@10 is used for evaluation, this test set would achieve a minimum detectable range of 0.20 for comparing $m = 10$ systems; also, this test set would be expected to make the confidence interval width of the difference between any systems be 0.20 or smaller;
- If nG@1 is used for evaluation, this test set would achieve a minimum detectable range of 0.20 for comparing $m = 5$ systems.

[*15] When $m = 2$, one-way ANOVA is equivalent to the unpaired t-test.

Table 4 Topic Set Size Design Results for STC $(\alpha, \beta) = (0.05, 0.20)$.

minD	$m = 2$	$m = 5$	$m = 10$	$m = 50$	$m = 100$
P+ ($\hat{\sigma}^2 = .0637$)					
0.05	391	604	794	1524	2056
0.10	**98**	152	199	382	515
0.15	44	68	**89**	170	229
0.20	25	39	50	**96**	129
nERR@10 ($\hat{\sigma}^2 = .0643$)					
0.05	395	609	802	1539	2075
0.10	**99**	153	201	385	519
0.15	45	68	**90**	172	231
0.20	26	39	51	**97**	130
nG@1 ($\hat{\sigma}^2 = .1515$)					
0.05	928	1434	1888	3625	4889
0.10	233	359	473	907	1223
0.15	104	160	211	403	544
0.20	59	**90**	119	227	306

In Table 4, the topic set sizes that correspond to the above discussions are shown in bold. Topic set size design can thus provide justifications for a particular decision on the number of topics included in a new test collection.

Previous work has shown that, from a statistical viewpoint, it is more economical to have many topics with a small number of judgments than to have a small number of topics with many judgments (e.g. [13], [14], [17], [20]). The STC task follows these recommendations and plans to rely on depth-10 pools. At the time of this writing, we have 15 teams that have signed up for the STC task; if each team submits five runs, we will have 75 runs in

total. The pool size will therefore be $75 * 10 = 750$ in the worst case (though this will in fact be around a few hundreds due to overlaps across runs); hence, if we have 100 test topics (posts), 75,000 comments will have to be assessed in the worst case. The STC organisers have enough budget to hire multiple assessors to judge each comment. We shall report on inter-assessor agreement in our STC overview paper in 2016.

5. Conclusions

In this study, we applied the ANOVA-based topic set size design technique of Sakai to determine the size of the test set for the NTCIR-12 STC task. Our main conclusion is to create 100 test topics, but what distinguishes our work from other tasks with similar topic set sizes is that we know what this topic set size means from a statistical viewpoint for each of our evaluation measures. We also demonstrated that, under the same set of statistical requirements, the topic set sizes required by nERR@10 and P^+ are more or less the same, while nG@1 requires more than twice as many topics. To our knowledge, our task is the first among all efforts at TREC-like evaluation conferences to actually create a new test collection by using this principled approach.

There are a few limitations to the present study. First, our training data labels were devised based on pooling only three runs, which probably means that they are highly incomplete and biased. Our six runs used for estimating the within-system variances of the three evaluation measures were evaluated using the incomplete training labels. The fundamental assumption behind the present study is that the estimates of the within-system variances ($\hat{\sigma}^2$'s) are of reasonable accuracy despite the above limitations. We shall verify whether our $\hat{\sigma}^2$'s are indeed reasonably accurate once we have collected the official STC runs from participants and have completed the contruction of the test data labels. Using the new topic-by-run matrices, where the rows represent 100 new topics and the columns represent the STC participants' runs, we will obtain more accurate estimates of the $\hat{\sigma}^2$ for each evaluation measure. Using these new estimates, we can decide on the topic set sizes for the *next* round of STC. We believe that, in this way, tasks should keep trying to improve the design of their test collections in terms of statistical reliability. Our hope is that the present effort will set a good example for other tasks at TREC-like evaluation conferences.

Acknowledgments The author would like to thank Lifeng Shang, Zhengdong Li and Hang Li of Huawei Noah's Ark Lab for their contributions to the present study as the NTCIR-12 STC organisers.

References

[1] Broder, A.: A Taxonomy of Web Search, *SIGIR Forum*, Vol. 36, No. 2, pp. 3–10 (2002).
[2] Chapelle, O., Ji, S., Liao, C., Velipasaoglu, E., Lai, L. and Wu, S.-L.: Intent-based Diversification of Web Search Results: Metrics and Algorithms, *Information Retrieval*, Vol. 14, No. 6, pp. 572–592 (2011).
[3] Ellis, P. D.: *The Essential Guide to Effect Sizes*, Cambridge University Press (2010).
[4] Higashinaka, R., Kawamae, N., Sadamitsu, K., Minami, Y., Meguro, T., Dohsaka, K. and Inagaki, H.: Building a Conversational Model from Two-Tweets, *Proceedings of IEEE ASRU 2011* (2011).
[5] Jafarpour, S. and Burges, C. J.: Filter, Rank and Transfer the Knowledge: Learning to Chat, Technical report, MSR-TR-2010-93 (2010).
[6] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM TOIS*, Vol. 20, No. 4, pp. 422–446 (2002).
[7] Lin, J. and Efron, M.: Overview of the TREC-2013 Microblog Track, *Proceedings of TREC 2013* (2014).
[8] Mitamura, T., Shima, H., Sakai, T., Kando, N., Mori, T., Takeda, K., Lin, C.-Y., Song, R., Lin, C.-J. and Lee, C.-W.: Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access, *Proceedings of NTCIR-8*, pp. 15–24 (2010).
[9] Nagata, Y.: *How to Design the Sample Size (in Japanese)*, Asakura Shoten (2003).
[10] Ritter, A., Cherry, C. and Dolan, W. B.: Data-Driven Response Generation in Social Media, *Proceedings of EMNLP 2011*, pp. 583–593 (2011).
[11] Sakai, T.: Bootstrap-Based Comparisons of IR Metrics for Finding One Relevant Document, *AIRS 2006 (LNCS 4182)*, pp. 374–389 (2006).
[12] Sakai, T.: Statistical Reform in Information Retrieval?, *SIGIR Forum*, Vol. 48, No. 1 (2014).
[13] Sakai, T.: *Information Access Evaluation Methodology: For the Progress of Search Engines (in Japanese)*, Coronasha (2015).
[14] Sakai, T.: Topic Set Size Design, *Information Retrieval Journal (submitted)* (2015).
[15] Sakai, T., Ishikawa, D., Kando, N., Seki, Y., Kuriyama, K. and Lin, C.-Y.: Using Graded-Relevance Metrics for Evaluating Community QA Answer Selection, *Proceedings of ACM WSDM 2011*, pp. 187–196 (2011).
[16] Sakai, T. and Robertson, S.: Modelling A User Population for Designing Information Retrieval Metrics, *Proceedings of EVIA 2008*, pp. 30–41 (2008).
[17] Sanderson, M. and Zobel, J.: Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability, *Proceedings of ACM SIGIR 2005*, pp. 162–169 (2005).
[18] Shibuki, H., Sakamoto, K., Kano, Y., Mitamura, T., Ishioroshi, M., Itakura, K., Wang, D., Mori, T. and Kando, N.: Overview of the NTCIR-11 QA-Lab Task, *Proceedings of NTCIR-11*, pp. 518–529 (2014).
[19] Voorhees, E. M. and Harman, D. K.(eds.): *TREC: Experiment and Evaluation in Information Retrieval*, The MIT Press (2005).
[20] Webber, W., Moffat, A. and Zobel, J.: Statistical Power in Retrieval Experimentation, *Proceedings of ACM CIKM 2008*, pp. 571–580 (2008).
[21] Weizenbaum, J.: ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine, *Communications of the ACM*, Vol. 9, No. 1 (1966).