Toshiba BRIDJE at NTCIR-6 CLIR: The Head/Lead Method and Graded Relevance Feedback

Tetsuya Sakai*† Makoto Koyama* Tatsuya Izuha*
Akira Kumano* Toshihiko Manabe* Tomoharu Kokubu*
*Toshiba Corporate R&D Center / †NewsWatch, Inc. (current affiliation)
sakai@newswatch.co.jp

Abstract

At NTCIR-6 CLIR, Toshiba participated in the Monolingual and Bilingual IR tasks covering three topic languages (Japanese, English and Chinese) and one document language (Japanese). For Stage 1 (which is the usual ad hoc task using the new NTCIR-6 topics), we submitted two DESCRIPTION runs and two TITLE runs for each topic language. Our first search strategy is Selective Sampling with Memory Resetting, and our second one is the Head/Lead method, which uses the Selective Sampling run as one of the components for data fusion. According to the Relaxed and Rigid Mean Average Precision statistics released by the organisers, we are the top performer in all six subtasks. For Stage 2 (which reused the NTCIR-3, 4 and 5 test collections), we repeated our two Stage 1 strategies in order to enable analysis across all four test collections. Moreover, we conducted some unofficial true relevance feedback experiments by exploiting the graded relevance data provided in the test collections. Our automatic run results show that the Head/Lead method slightly but consistently improves performance, while our unofficial "interactive" run results suggest that graded-relevance metrics favour graded relevance feedback while Average Precision favours binary relevance feedback. In addition, our significance tests suggest that the NTCIR-6 Japanese test collection is "harder" than previous collections.
Keywords: *Selective Sampling, Head/Lead method, Graded Relevance Feedback, Q-measure, Geometric Mean, Bootstrap Hypothesis Test.*

1 Introduction

At NTCIR-6 CLIR, Toshiba participated in the Monolingual and Bilingual IR tasks covering three topic languages (Japanese, English and Chinese) and one document language (Japanese). For Stage 1 (which is the usual ad hoc task using the new NTCIR-6 topics), we submitted two DESCRIPTION runs and two TITLE runs for each topic language. Our first search strategy is *Selective Sampling with Memory Resetting* [7], and our second one is the Head/Lead method, which uses the Selective Sampling run as one of the components for data fusion. According to the Relaxed and Rigid Mean Average Precision (MAP) statistics released by the organisers, we are the top performer in all six subtasks. For Stage 2 (which reused the NTCIR-3, 4 and 5 test collections), we repeated our two Stage 1 strategies in order to enable analysis across all four test collections. Moreover, we conducted some unofficial true relevance feedback experiments by exploiting the graded relevance data provided in the test collections. Our automatic run results show that the Head/Lead method slightly but consistently improves performance, while our unofficial "interactive" run results suggest that graded-relevance metrics favour *graded relevance feedback* (e.g. [1, 16]) while Average Precision favours binary relevance feedback. In addition, our significance tests suggest that the NTCIR-6 Japanese test collection is "harder" than previous collections.

Tables 1-3 show our automatic run results at NTCIR-6 in terms of Relaxed MAP, Mean Q-measure (MQ) and Geometric Mean Q-measure (GMQ) using the default gain values 3:2:1 [8, 9]. The first column of each table shows the run names we shall use in this paper. For example, a run with the JJ-D prefix is a Japanese monolingual run using the DESCRIPTION field. Whereas, the suffixes noPRF, PRF, SSR and H/L stand for no pseudo-relevance feedback, pseudo-relevance feedback, selective sampling (with memory resetting) and the Head/Lead method, respectively. The second column of each table shows the official run names, if applicable. For example, TSB-J-J-D-01 is our first Japanese monolingual run for the NTCIR-6 test collection at Stage 1, and TSB-J-J-D-01-N3 is our first Japanese monolingual run for the NTCIR-3 test collection at Stage 2 [4]. **Boldface** indicates which of our two strategies are better on average.

Since we have ample evidence that Q-measure is a reliable metric that can handle graded relevance [9,

Table 1. TSB's automatic run results at NTCIR-6 (Relaxed MAP).

Name	Official Name	NTCIR-3	NTCIR-4	NTCIR-5	NTCIR-6
(a) Monolingual DESCRIPTION runs					
JJ-D-noPRF	-	.4085	.4072	.3791	.3041
JJ-D-PRF	-	.4587	.4974	.4775	.4055
JJ-D-SSR	TSB-J-J-D-01{-N3,-N4,-N5,}	.4613	.5031	.4792	.4090
JJ-D-H/L	TSB-J-J-D-02{-N3,-N4,-N5,}	**.4702**	**.5082**	**.4911**	**.4138**
(b) Monolingual TITLE runs					
JJ-T-noPRF	-	.3973	.4050	.3684	.3258
JJ-T-PRF	-	.4668	.5037	.4840	.4326
JJ-T-SSR	TSB-J-J-T-03{-N3,-N4,-N5,}	.4580	.5045	.4941	.4375
JJ-T-H/L	TSB-J-J-T-04{-N3,-N4,-N5,}	**.4594**	**.5069**	**.5046**	**.4393**
(c) English-Japanese DESCRIPTION runs					
EJ-D-noPRF	-	.3696	.3033	.3161	.2572
EJ-D-PRF	-	.4432	.4491	.4019	.3600
EJ-D-SSR	TSB-E-J-D-01{-N3,-N4,-N5,}	.4329	.4449	.4085	.3665
EJ-D-H/L	TSB-E-J-D-02{-N3,-N4,-N5,}	**.4381**	**.4512**	**.4198**	**.3686**
(d) English-Japanese TITLE runs					
EJ-T-noPRF	-	.2990	.2960	.2917	.2537
EJ-T-PRF	-	.3727	.4676	.3957	.3443
EJ-T-SSR	TSB-E-J-T-03{-N3,-N4,-N5,}	.3746	.4552	.4025	**.3588**
EJ-T-H/L	TSB-E-J-T-04{-N3,-N4,-N5,}	**.3806**	**.4610**	**.4169**	.3576
(e) Chinese-Japanese DESCRIPTION runs					
CJ-D-noPRF	-	.3555	.3124	.2879	.2571
CJ-D-PRF	-	.4299	.4214	.4028	.3645
CJ-D-SSR	TSB-C-J-D-01{-N3,-N4,-N5,}	.4213	.4129	.4005	.3650
CJ-D-H/L	TSB-C-J-D-02{-N3,-N4,-N5,}	**.4279**	**.4193**	**.4187**	**.3713**
(f) Chinese-Japanese TITLE runs					
CJ-T-noPRF	-	.3022	.3357	.2987	.2804
CJ-T-PRF	-	.3829	.4493	.4305	.3848
CJ-T-SSR	TSB-C-J-T-03{-N3,-N4,-N5,}	.3692	**.4450**	.4238	.3820
CJ-T-H/L	TSB-C-J-T-04{-N3,-N4,-N5,}	**.3850**	.4416	**.4303**	**.3849**

13, 11][1], we use MQ as our primary summary performance statistic. Throughout this paper, statistical significance is discussed based on the MQ values only, using the two-tailed *paired bootstrap hypothesis test* as described in [9] by default.

Robertson [5] has discussed the benefit of using the geometric mean in addition to the arithmetic mean for the purpose of obtaining a summary performance value: Hence the GMQ values are shown in Table 3. As we discussed in [8], this provides a good preview of what is happening per-topic. For example, Table 4 shows the *relative* performances of our H/L-based cross-language runs as compared to the corresponding monolingual runs, computed based on the absolute values from Tables 2 and 3. It can be observed that the GMQ-based relative performances are much lower than the MQ-based ones. For example, for NTCIR-3, CJ-T-H/L is 82.7% of the corresponding monolingual run JJ-T-H/L according to MQ, but is only 46.3% of JJ-T-H/L according to GMQ. This uncovers the fact that the failure of search request translation is very serious for some topics. However, we shall hereafter focus on the new Head/Lead method and Graded Relevance Feedback for *monolingual* IR.

The remainder of this paper is organised as follows. Section 2 briefly describes our search strategies. Section 3 discusses our official automatic runs, namely, those based on Selective Sampling and the Head/Lead method. In addition, it also compares the four NTCIR test collections using our Head/Lead runs. Section 4 discusses our unofficial graded relevance feedback runs which make use of the test collection "qrels" for query expansion and are therefore categorised as "interactive" runs, although no user effort was actually spent for these experiments. Finally, Section 5 concludes this paper.

2 Search Strategies

2.1 BRIDJE

For performing Japanese document retrieval, we used the retriever component of the BRIDJE system [6] which indexes documents and processes topics using morphological analysis. By default, BRIDJE uses Okapi/BM25 term weighting [15] and Pseudo-Relevance Feedback (PRF) based on the *offer weight* (ow) [12, 15] for term selection.

English and Chinese topics were translated into Japanese using the Toshiba Machine Translation (MT) system. As our focus for this year was on monolingual IR, we took the "black-box" MT approach as opposed to *partial disambiguation* [6, 8] which preserves two or more translations per source query term[2].

[1] Q-measure has also proved to be a reliable evaluation metric for tasks other than traditional document retrieval: Question Answering [12] and XML retrieval at INEX [3].

[2] It should be noted that all CLIR topics used at NTCIR-6 CLIR Stages 1 and 2 were "known", in the sense that participating systems had already encountered them during the previous NTCIR rounds. (See Figure 3, which we shall discuss later.) Hence the CLIR performances reported at NTCIR-6 may not necessarily be representative of a CLIR situation in which unknown incoming search requests need to be translated.

Table 2. TSB's automatic run results at NTCIR-6 (Mean Q-measure).

Name	Official Name	NTCIR-3	NTCIR-4	NTCIR-5	NTCIR-6
(a) Monolingual DESCRIPTION runs					
JJ-D-noPRF	-	.4367	.4217	.4132	.3377
JJ-D-PRF	-	.4881	.5098	.5022	.4341
JJ-D-SSR	TSB-J-J-D-01{-N3,-N4,-N5,}	.4940	.5160	.5037	.4353
JJ-D-H/L	TSB-J-J-D-02{-N3,-N4,-N5,}	**.5010**	**.5206**$*$	**.5147**$**$	**.4402**
(b) Monolingual TITLE runs					
JJ-T-noPRF	-	.4400	.4265	.4067	.3682
JJ-T-PRF	-	.4895	.5164	.5096	.4600
JJ-T-SSR	TSB-J-J-T-03{-N3,-N4,-N5,}	.4827	.5187	.5223	.4635
JJ-T-H/L	TSB-J-J-T-04{-N3,-N4,-N5,}	**.4845**	**.5199**	**.5321**$*$	**.4648**
(c) English-Japanese DESCRIPTION runs					
EJ-D-noPRF	-	.4054	.3554	.3501	.2917
EJ-D-PRF	-	.4755	.4591	.4259	.3935
EJ-D-SSR	TSB-E-J-D-01{-N3,-N4,-N5,}	.4651	.4582	.4317	.3992
EJ-D-H/L	TSB-E-J-D-02{-N3,-N4,-N5,}	**.4693**	**.4633**$*$	**.4424**$**$	**.4021**
(d) English-Japanese TITLE runs					
EJ-T-noPRF	-	.3390	.3686	.3288	.2901
EJ-T-PRF	-	.4007	.4796	.4187	.3708
EJ-T-SSR	TSB-E-J-T-03{-N3,-N4,-N5,}	.4039	.4706	.4273	**.3869**
EJ-T-H/L	TSB-E-J-T-04{-N3,-N4,-N5,}	**.4094**	**.4739**	**.4405**$**$	.3862
(e) Chinese-Japanese DESCRIPTION runs					
CJ-D-noPRF	-	.3745	.3296	.3222	.2924
CJ-D-PRF	-	.4522	.4393	.4285	.3966
CJ-D-SSR	TSB-C-J-D-01{-N3,-N4,-N5,}	.4474	.4312	.4270	.3958
CJ-D-H/L	TSB-C-J-D-02{-N3,-N4,-N5,}	**.4527**	**.4375**$*$	**.4456**$**$	**.4033**
(f) Chinese-Japanese TITLE runs					
CJ-T-noPRF	-	.3315	.3600	.3401	.3211
CJ-T-PRF	-	.3997	.4624	.4566	.4129
CJ-T-SSR	TSB-C-J-T-03{-N3,-N4,-N5,}	.3897	**.4609**	.4502	.4084
CJ-T-H/L	TSB-C-J-T-04{-N3,-N4,-N5,}	**.4005**$*$	.4581	**.4576**$**$	**.4113**

Statistically significant differences between SSR and H/L are indicated by $*$ ($\alpha = 0.05$) and $**$ ($\alpha = 0.01$).

Table 3. TSB's automatic run results at NTCIR-6 (Geometric Mean Q-measure).

Name	Official Name	NTCIR-3	NTCIR-4	NTCIR-5	NTCIR-6
(a) Monolingual DESCRIPTION runs					
JJ-D-noPRF	-	.3555	.3253	.3412	.2366
JJ-D-PRF	-	.3714	.4056	.3985	.3202
JJ-D-SSR	TSB-J-J-D-01{-N3,-N4,-N5,}	.3871	.4126	.4008	.3256
JJ-D-H/L	TSB-J-J-D-02{-N3,-N4,-N5,}	**.3934**	**.4175**	**.4122**	**.3294**
(b) Monolingual TITLE runs					
JJ-T-noPRF	-	.2835	.3484	.3223	.2693
JJ-T-PRF	-	.3323	.4442	.4207	.3601
JJ-T-SSR	TSB-J-J-T-03{-N3,-N4,-N5,}	.3226	.4500	.4397	.3582
JJ-T-H/L	TSB-J-J-T-04{-N3,-N4,-N5,}	**.3257**	**.4523**	**.4469**	**.3625**
(c) English-Japanese DESCRIPTION runs					
EJ-D-noPRF	-	.2779	.2406	.2336	.1845
EJ-D-PRF	-	.2854	.3502	.2114	.2617
EJ-D-SSR	TSB-E-J-D-01{-N3,-N4,-N5,}	**.2820**	.3487	.2397	.2655
EJ-D-H/L	TSB-E-J-D-02{-N3,-N4,-N5,}	.2819	**.3504**	**.2500**	**.2662**
(d) English-Japanese TITLE runs					
EJ-T-noPRF	-	.1434	.2443	.1644	.1751
EJ-T-PRF	-	.1529	.3446	.1933	.2032
EJ-T-SSR	TSB-E-J-T-03{-N3,-N4,-N5,}	.1614	.3345	.2172	.2092
EJ-T-H/L	TSB-E-J-T-04{-N3,-N4,-N5,}	**.1658**	**.3385**	**.2225**	**.2132**
(e) Chinese-Japanese DESCRIPTION runs					
CJ-D-noPRF	-	.2351	.2203	.2270	.1935
CJ-D-PRF	-	.2842	.3162	.2732	.2480
CJ-D-SSR	TSB-C-J-D-01{-N3,-N4,-N5,}	.2818	.3030	.2780	.2551
CJ-D-H/L	TSB-C-J-D-02{-N3,-N4,-N5,}	**.2881**	**.3099**	**.2951**	**.2634**
(f) Chinese-Japanese TITLE runs					
CJ-T-noPRF	-	.1141	.2343	.2408	.2083
CJ-T-PRF	-	.1582	.3482	.3331	.2762
CJ-T-SSR	TSB-C-J-T-03{-N3,-N4,-N5,}	.1413	.3468	.3341	.2740
CJ-T-H/L	TSB-C-J-T-04{-N3,-N4,-N5,}	**.1508**	**.3496**	**.3438**	**.2747**

Table 4. Relative CLIR performances compared to monolingual ones.

Mean Q-measure	NTCIR-3	NTCIR-4	NTCIR-5	NTCIR-6
EJ-D-H/L	93.7%	89.0%	86.0%	91.3%
EJ-T-H/L	84.5%	91.2%	82.8%	83.1%
CJ-D-H/L	90.4%	84.0%	86.6%	91.6%
CJ-T-H/L	82.7%	88.1%	86.0%	88.5%
Geometric Mean Q-measure	NTCIR-3	NTCIR-4	NTCIR-5	NTCIR-6
EJ-D-H/L	71.7%	83.9%	60.7%	80.8%
EJ-T-H/L	50.9%	74.8%	49.8%	58.8%
CJ-D-H/L	73.2%	74.2%	71.6%	80.0%
CJ-T-H/L	46.3%	77.3%	76.9%	75.8%

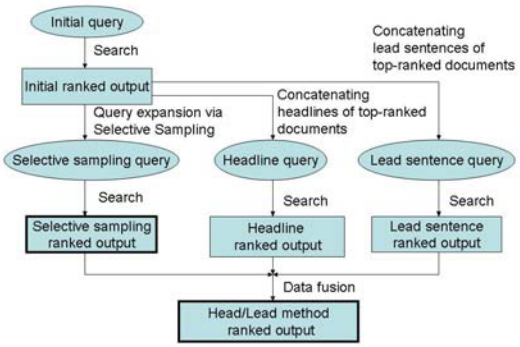

Figure 1. The Head/Lead method.

2.2 Selective Sampling

Our first official run strategy is Selective Sampling (with memory resetting) proposed in [7]. The only difference between this method and default PRF is the way pseudo-relevant documents are selected: From the initial ranked output (i.e., noPRF), default PRF takes the top $P = 20$ documents; Selective Sampling takes between $P_{min} = 5$ and $P_{max} = 20$ documents from the top $P_{scope} = 100$ documents, possibly skipping "similar" documents. Our previous work showed that Selective Sampling outperforms PRF at least as often as PRF outperforms Selective Sampling on a per-topic basis, and that the two methods are comparable in terms of average performance. Encouraged by these results, we chose Selective Sampling instead of default PRF as our first official run strategy. Both PRF and Selective Sampling runs used $T = 30$ expansion terms.

2.3 The Head/Lead Method

Our second official run strategy is a new method called the The Head/Lead method. Since it is known that data fusion is a promising technique for enhancing performance, we designed one such method for newspaper articles. As shown in Figure 1, the Head/Lead method creates three queries independently using the initial query and the initial ranked output: A selective sampling query (i.e., an expanded query), a *headline query* created by concatenating the headlines of top 10 initially retrieved documents, and a *lead sentence query* created by concatenating the first sentences of top 5 initially retrieved documents. Three different representations of the same search request are thus obtained. A search is performed using each query, and finally the three ranked lists are merged, by taking a weighted average of the document scores [8]. Through some tuning experiments using the NTCIR-3, 4 and 5 test collections, the weight ratio was set to 7:2:1, thus giving the highest weight to the Selective Sampling output. Note that the Head/Lead runs should primarily be compared with the raw Selective Sampling run rather than the default PRF run.

2.4 Graded Relevance Feedback

In addition to our official automatic run experiments, we conducted some true relevance feedback experiments using the "qrel" files in order to address the following questions:

(a) How does *graded relevance feedback* (e.g. [1, 16]), which assumes that the user provides the *relevance levels* of some retrieved documents to the system, compare to traditional binary relevance feedback? Would graded-relevance metrics (MQ and GMQ) agree with the binary MAP as to which kind of feedback is better?

(b) How does relevance feedback using all relevant documents from the top 20 initially retrieved documents compare to pseudo-relevance feedback using *all* of the top 20 documents?

(c) What is the upperbound of binary and graded relevance feedback?

Our graded relevance feedback algorithm is a natural extension of binary relevance feedback based on the *offer weight* [15]. Let $gw(\mathcal{L})$ denote the *grade weight* for an $\mathcal{L}$-relevant document ($\mathcal{L} \in \{S, A, B\}$), which reflects how an $\mathcal{L}$-relevant document fed back to the system should contribute to query expansion. By default, we let $gw(S) = 3, gw(A) = 2, gw(B) = 1$. Let $rw(t)$ denote the Okapi *relevance weight* [15] for a term t, and let $r_{\mathcal{L}}(t)$ denote the number of known $\mathcal{L}$-relevant documents containing t. Then, our term selection criterion, or *graded offer weight*, is defined as:

$$gow(t) = rw(t) * \sum_{\mathcal{L}} gw(\mathcal{L}) * r_{\mathcal{L}}(t) \qquad (1)$$

It is clear that, when $gw(S) = gw(A) = gw(B) = 1$, graded offer weight reduces to the traditional offer weight used in binary relevanced feedback. Note also that our graded relevance feedback algorithm still relies on the traditional relevance weight, which is based purely on the binary-relevance probabilistic model.

Our first set of relevance feedback runs, which we call collectively as top20RF, used relevant documents in the top 20 initially retrieved documents for each topic. That is, we assumed that the user examines top 20 documents exhaustively for every topic and identifies highly relevant, relevant and partially relevant documents. Comparing these runs with our automatic PRF runs, which treated *all* of the top 20 documents as relevant, should provide an answer to Question (b) posed above.

Our second set of relevance feedback runs, which we call collectively as allRF, was designed to answer Question (c) posed above. Thus, in order to provide an upperbound of our relevance feedback algorithm, these runs used *all* relevant documents, not just initially retrieved ones. Note that they do not have any

practical significance, since a system that is given all relevant documents should just list them up in order to achieve the best-possible performance.

With both top20RF and allRF, we tried three different grade weight ratios: $gw(S) : gw(A) : gw(B) = 1 : 1 : 1$ (representing binary relevance feedback), $1 : 1 : 0$ (ignoring partially relevant documents) and $3 : 2 : 1$. Comparing the effect of these different ratios interms of MAP, MQ and GMQ should provide some answers to Question (a) posed above. All of these runs used $T = 30$ new terms just like our automatic runs.

3 Official Automatic Run Results: The Head/Lead Method

As mentioned earlier, our official runs results are included in Tables 1-3. As with previous NTCIR rounds, the feedback runs (PRF, SSR, H/L) are substantially better than the corresponding no-feedback runs (no-PRF); SSR is as effective as PRF on average, even with the new NTCIR-6 data. The remainder of this section focusses on our best official strategy: H/L.

3.1 Head/Lead vs Selective Sampling

We have conducted significance tests for the differences in MQ between SSR and H/L, and the results are indicated by "$*$" ($\alpha = 0.05$) and "$**$" ($\alpha = 0.01$) in Table 2. The advantage of H/L over SSR is clear for the NTCIR-5 data, but the differences are not statistically significant for the NTCIR-6 data, which is our test data. However, even for this new data set, H/L slightly outperforms SSR on average in most cases. In summary, the effect of the Head/Lead method is small, but the approach of utilising newspaper headlines and lead sentences is probably a step in the right direction.

Figure 2 visualises the per-topic Q-measure values of our monolingual H/L runs for the NTCIR-6 test collection. It can be observed that BRIDJE fails completely for the Topic 019 DESCRIPTION (Q-measure=.0005), but we should first note that the topic set has a bug here: The English DESCRIPTION for this topic is "... international incidents *at sea*, involving more than one country", but the Japanese DESCRIPTION does not mention the sea at all; Whereas, the Japanese TITLE does mention the sea and therefore does better (Q-measure=.1309). However, with or without the bug, the topic *is* challenging, in that it calls for two or more countries without naming them specifically. A possible approach would be to perform named entity recognition on indexed/retrieved documents and to count the number of distinct instances tagged with COUNTRY: We have tried some preliminary query-specific approaches such as this but without consistent success.

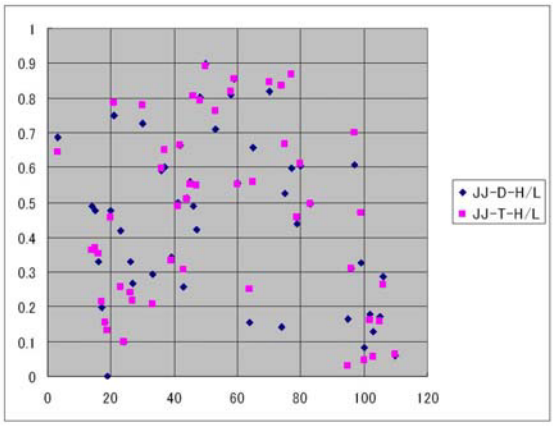

Figure 2. Per-topic Q-measure values of the monolingual H/L runs (NTCIR-6).

3.2 Collection Comparisons

We now discuss the "hardness" of the Japanese NTCIR test collections using the JJ-D-H/L and JJ-T-H/L runs. This is possible because we used exactly the same search strategy with each test collection. Sakai [10] reported on a similar analysis but he used the NTCIR-3 and NTCIR-5 test collections only.

We first conduct a pairwise comparison of the four test collections, by comparing the monolingual H/L performances shown in Table 2(a) using the two-tailed *unpaired* bootstrap hypothesis test [9, 10]. For example, we set up a null hypothesis that the NTCIR-3 performance values and the NTCIR-4 performance values come from an identical distribution. Table 5 shows the achieved significance levels obtained for each pair of test collections. It can be observed that none of the cross-collection differences is statistically significant according to the unpaired test.

In the case of NTCIR CLIR, however, more direct collection comparisons are possible. As illustrated in Figure 3, the 50 NTCIR-6 Japanese topics originate from NTCIR-3 and NTCIR-4. To be more precise, the first 31 NTCIR-6 Japanese topics are from NTCIR-4, and for these topics, the NTCIR-6 topic IDs are the same as the original NTCIR-4 topic IDs; the remaining 19 NTCIR-6 Japanese topics are from NTCIR-3, and for each of these topics, the NTCIR-6 topic ID can be obtained by adding 60 to the original NTCIR-3 topic ID. For example, NTCIR-6 Topic 064 is exactly NTCIR-3 Topic 004. Thus, let Q_{N3}, Q_{N4} and Q_{N6} represent the Japanese topic sets from NTCIR-3, -4 and -6, respectively, where $|Q_{N3}| = 42$, $|Q_{N4}| = 55$ and $|Q_{N6}| = 50$. Then we can directly compare the NTCIR-6 results with the NTCIR-3 and NTCIR-4 results using *paired* significance tests, by considering the topic sets $Q_{N3} \cap Q_{N6}$ and $Q_{N4} \cap Q_{N6}$, where $|Q_{N3} \cap Q_{N6}| = 19$ and $|Q_{N4} \cap Q_{N6}| = 31$.

Table 6 shows the monolingual H/L performances averaged over $Q_{N3} \cap Q_{N6}$ and $Q_{N4} \cap Q_{N6}$, together

Table 5. Achieved siginificance levels (unpaired bootstrap test).

JJ-D-H/L	NTCIR-4	NTCIR-5	NTCIR-6
NTCIR-3	$p = .651$	$p = .785$	$p = .211$
NTCIR-4	-	$p = .886$	$p = .069$
NTCIR-5	-	-	$p = .124$
JJ-T-H/L	NTCIR-4	NTCIR-5	NTCIR-6
NTCIR-3	$p = .483$	$p = .391$	$p = .730$
NTCIR-4	-	$p = .758$	$p = .243$
NTCIR-5	-	-	$p = .146$

with the results of paired bootstrap hypothesis tests. For example, the JJ-D-H/L strategy run against the NTCIR-3 collection and averaged over $Q_{N3} \cap Q_{N6}$ achieves .5302 in MQ, while the same strategy run against the NTCIR-6 collection and averaged over the same topic set achieves .3552, and this difference is statistically very highly significant. Note that the p-values for the differences between NTCIR-4 and NTCIR-6 are also very low, though not significant. Thus, according to these paired tests, which have much higher power than the aforementioned unpaired tests [9, 10], the NTCIR-6 Japanese test collection appears to be "harder" than previous collections.

One possible hypothesis for explaining the fact that the NTCIR-6 performances are considerably lower than those with the older collections is that the NTCIR-6 relevance data may be more *incomplete* [2] than others[3]. It is known that standard IR metrics computed based on incomplete relevance data *underestimate* system performances. Yilmaz and Aslam [17] and Sakai [14] independently showed that this problem in an incomplete relevance environment can be remedied using *condensed lists*, obtained by *removing all unjudged documents* from the original ranked list prior to applying a standard IR metric. This approach is actualy more robust to incompleteness than *bpref* [2], which was designed specifically for handling the incompleteness problem. Following Sakai [14], we couple Q-measure with condensed lists, and the new metric will be referred to as Q′. Moreover, the Mean of Q′ values across a topic set will be referred to as MQ′. If our hypothesis is correct, then the performance gap between the NTCIR-6 results and the NTCIR-3/4 ones may be smaller in terms of MQ′, since the absolute MQ′ values are much more robust to incompleteness than the absolute MQ values which are known to diminish quickly as the test collection becomes more and more incomplete [14].

Table 7 shows the MQ′ values for our monolingual H/L runs, using the full topic set for each data set. It can be observed that the values are somewhat higher than the corresponding MQ values shown in Table 2(a)(b), but the gaps between the NTCIR-6 results and others still persist. Table 8 repeats the paired-test comparisons between test collection pairs but this time using MQ′ instead of MQ. Again, the p-values are not much different from those in Table 6. In summary, the performance gaps between NTCIR-6 and the other collections exist even in terms of MQ′, which is relatively robust to incompleteness. Thus, our hypthesis that NTCIR-6 relevance data is more incomplete than others is not supported. Perhaps the Organisers can shed light on this issue.

[3] Any test collections constructed through *pooling* are inherently incomplete in that not all documents in the collections have been judged for relevance.

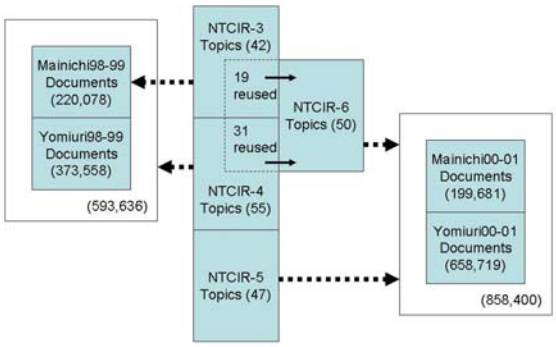

Figure 3. The NTCIR-3,4,5,6 Japanese collections.

Table 6. MQ values and achieved siginificance levels (paired bootstrap test).

JJ-D-H/L	NTCIR-3	NTCIR-4	NTCIR-6	ASL
$Q_{N3} \cap Q_{N6}$	.5302	-	.3552	$p = .000 **$
$Q_{N4} \cap Q_{N6}$	-	.5444	.4923	$p = .054$
JJ-T-H/L	NTCIR-3	NTCIR-4	NTCIR-6	ASL
$Q_{N3} \cap Q_{N6}$	.5023	-	.4128	$p = .004 **$
$Q_{N4} \cap Q_{N6}$	-	.5479	.4966	$p = .075$

Table 7. MQ′ values based on condensed lists (*cf.* Table 2(a)(b)).

	NTCIR-3	NTCIR-4	NTCIR-5	NTCIR-6
JJ-D-H/L	.5145	.5340	.5225	.4536
JJ-T-H/L	.4993	.5333	.5399	.4776

Table 8. MQ′ values and achieved siginificance levels (paired bootstrap test).

JJ-D-H/L	NTCIR-3	NTCIR-4	NTCIR-6	ASL
$Q_{N3} \cap Q_{N6}$	.5464	-	.3688	$p = .000 **$
$Q_{N4} \cap Q_{N6}$	-	.5586	.5056	$p = .053$
JJ-T-H/L	NTCIR-3	NTCIR-4	NTCIR-6	ASL
$Q_{N3} \cap Q_{N6}$	.5194	-	.4221	$p = .026*$
$Q_{N4} \cap Q_{N6}$	-	.5628	.5116	$p = .058$

4 Unofficial "Interactive" Run Results: Graded Relevance Feedback

Table 9 summarises the results of our unofficial true relevance feedback (RF) experiments. For example, JJ-D-top20RF1:1:1 is a DESCRIPTION run using binary RF based on all relevant documents from the top 20 initially retrieved set, while JJ-D-allRF1:1:1 is the corresponding run using *all* known relevant documents for RF. The PRF performances from Tables 1-

Table 9. Unofficial "interactive" results: graded relevance feedback.

Metric	Name	NTCIR-3	NTCIR-4	NTCIR-5	NTCIR-6
(a) Relaxed MAP	(JJ-D-PRF)	.4587	.4974	.4775	.4055
	JJ-D-top20RF1:1:1	.5492	**.5280**	.5526	**.4527**
	JJ-D-top20RF1:1:0	**.5540**	.5265	**.5533**	.4503
	JJ-D-top20RF3:1:0	.5472	.5267	.5475	.4478
	JJ-D-allRF1:1:1	**.6732**	**.5976**	**.6383**	**.5512**
	JJ-D-allRF1:1:0	.6718	.5935	.6347	.5487
	JJ-D-allRF3:1:0	.6671	.5933	.6336	.5473
	(JJ-T-PRF)	.4668	.5037	.4840	.4326
	JJ-T-top20RF1:1:1	**.5508**	**.5299**	.5630	**.4724**
	JJ-T-top20RF1:1:0	.5505	.5272	**.5701**	.4664
	JJ-T-top20RF3:1:0	.5467	.5275	.5629	.4648
	JJ-T-allRF1:1:1	**.6623**	**.5920**	**.6550**	**.5510**
	JJ-T-allRF1:1:0	.6538	.5893	.6480	.5477
	JJ-T-allRF3:1:0	.6495	.5885	.6453	.5459
(b) Mean Q-measure Statistically significant differences between top20RF/allRF and PRF are indicated by $\star$ ($\alpha = 0.05$) and $\star\star$ ($\alpha = 0.01$). Those between 1:1:0/3:1:0 (graded relevance feedback) and 1:1:1 (binary relevance feedback) are indicated by † ($\alpha = 0.05$). Each allRF run is significantly better than top20RF at $\alpha = 0.01$.	(JJ-D-PRF)	.4881	.5098	.5022	.4341
	JJ-D-top20RF1:1:1	.5683$\star\star$	.5294	.5678$\star\star$	.4691$\star$
	JJ-D-top20RF1:1:0	**.5883**$\star\star$	.5336	**.5749**$\star\star$	**.4732**$\star\star$
	JJ-D-top20RF3:1:0	.5819$\star\star$	**.5365**$\star$	.5735$\star\star$	.4718$\star\star$
	JJ-D-allRF1:1:1	.6930	.6000	.6496	.5730
	JJ-D-allRF1:1:0	**.7045**	.6052	.6567	.5786
	JJ-D-allRF3:1:0	.7008	**.6082**	**.6621**	**.5789**
	(JJ-T-PRF)	.4895	.5164	.5096	.4600
	JJ-T-top20RF1:1:1	.5637$\star$	.5330	.5800$\star\star$	.4897$\star$
	JJ-T-top20RF1:1:0	**.5769**$\star\star$	.5360	**.5937**$\star\star$†	**.4916**$\star$
	JJ-T-top20RF3:1:0	.5747$\star\star$	**.5387**	.5912$\star\star$	.4910$\star$
	JJ-T-allRF1:1:1	.6781	.5962	.6622	.5729
	JJ-T-allRF1:1:0	**.6834**	.6016	.6672	**.5789**
	JJ-T-allRF3:1:0	.6810	**.6055**†	**.6708**	.5783
(c) Geometric Mean Q-measure	(JJ-D-PRF)	.3714	.4056	.3985	.3202
	JJ-D-top20RF1:1:1	.4859	.4375	.5044	.3872
	JJ-D-top20RF1:1:0	**.5405**	.4386	**.5143**	**.3892**
	JJ-D-top20RF3:1:0	.5317	**.4414**	.5132	.3846
	JJ-D-allRF1:1:1	.6779	.5698	.6288	.5254
	JJ-D-allRF1:1:0	**.6878**	.5733	**.6343**	**.5329**
	JJ-D-allRF3:1:0	.6840	**.5766**	.6333	.5327
	(JJ-T-PRF)	.3323	.4442	.4207	.3601
	JJ-T-top20RF1:1:1	.4532	.4683	.4838	.3928
	JJ-T-top20RF1:1:0	**.5219**	.4679	**.5468**	**.3935**
	JJ-T-top20RF3:1:0	.5154	**.4715**	.5442	.3921
	JJ-T-allRF1:1:1	.6564	.5631	.6400	.5251
	JJ-T-allRF1:1:0	**.6605**	.5672	**.6433**	**.5309**
	JJ-T-allRF3:1:0	.6567	**.5711**	.6403	.5301

3 have been duplicated here for comparison. Paired bootstrap hypothesis test results in terms of MQ are shown also: Significant differences betweem an RF run and a PRF run is indicated by $\star$ ($\alpha = 0.05$) and $\star\star$ ($\alpha = 0.01$); Those between a graded RF run and the corresponding binary RF run (e.g. JJ-T-allRF3:1:0 vs. JJ-T-allRF1:1:1) are indicated by † ($\alpha = 0.05$); Although not indicated explicitly in the table, each allRF run is significantly better than the corresponding top20RF run at $\alpha = 0.05$. Our findings are:

- Not surprisingly, most top20RF runs are significantly better than the corresponding PRF run, as the abundance of "$\star$"'s indicates. However, the PRF performances are quite impressive compared to true RF especially for the NTCIR-4 and -6 data: For example, JJ-D-top20RF1:1:1 for NTCIR-4 achieves .5294 in MQ, while the corresponding PRF performance is .5098, and the difference between the two is not statistically significant. Similarly, although there is a statistically significant difference between JJ-D-top20RF1:1:1 for NTCIR-6 and the corresponding PRF run, the absolute difference in MQ between the two is small and may be *practically* insignificant (.4691 vs .4341). It is really remarkable that a method that does not consult the user at all can boost the IR performance so much, although this has been known since the early 1990's.
- Not surprisingly, all allRF runs substantially outperform the corresponding top20RF runs. Hence either (a) Using more relevant documents for feedback is better; or (b) Using *unretrieved* relevant documents (together with retrieved ones) for feedback is better than using only *retrieved* relevant documents. Although investigating the contributions of (a) and (b) is beyond the scope of this paper, we suspect that (b) has a larger impact than (a).
- Interestingly, although most of the differences between graded RF and binary RF are not statistically significant, MAP and (G)MQ tend to disagree as to which of the two RF strategies is better on average. MAP, a metric based on binary relevance, tends to prefer binary RF (runs labelled with 1:1:1): See the boldface values in Table 9(a). MQ and GMQ, which utilise graded relevance assessments, tend to

prefer graded RF (runs labelled with 1:1:0 or 3:1:0): See the boldface values in Table 9(b) and (c). Moreover, the GMQ values of the graded top20RF runs for NTCIR-3 are much higher than those of binary top20RF: For example, for NTCIR-3, JJ-T-top20RF1:1:0 achieves .5219 in GMQ, while JJ-T-top20RF1:1:1 achieves only .4532. Similarly, for NTCIR-5, JJ-T-top20RF1:1:0 achieves .5468 in GMQ, while JJ-T-top20RF1:1:1 achieves only .4838. Note also that JJ-T-top20RF1:1:0 for NTCIR-5 is significantly better than JJ-T-top20RF1:1:1 in Table 9(b). These results suggest not only that graded relevance feedback may be worthwhile if the user can provide graded relevance assessments to some documents, but also that *IR systems that utilise graded relevance assessments should be evaluated using IR metrics that utilise graded relevance*.

5 Conclusions

This paper reported on our NTCIR-6 CLIR experiments involving four (NTCIR-3,4,5 and 6) test collections. We are the official top performer in all of the six ad hoc subtasks in which we participated (Japanese monolingual, English-Japanese and Chinese-Japanese IR using either DESCRIPTION or TITLE).

Our official automatic run results show that the Head/Lead method, which involves data fusion of three independently created ranked lists, slightly but consistently improves performance. Although this is by no means a breakthrough, we believe that our method of utilising the headlines and lead sentences for retrieval of newspaper articles is a small step in the right direction. It is also clear now that Selective Sampling is at least as effective as traditional PRF, and that taking the top P documents is not necessarily the best choice, which confirms a finding in [7]. We also conducted some significance tests which suggest that the NTCIR-6 test collection is "harder" than others.

Our unofficial true RF results show that (i) True RF using relevant documents ranked within top 20 outperforms PRF using all of the 20 documents, but the difference between the two is relatively small in some cases; (ii) True RF using all known relevant documents outperforms true RF using relevant documents ranked within top 20, either because it uses more relevant documents, or because it uses unretrieved relevant documents, or both; (iii) MAP favours binary RF, while MQ and GMQ favour graded RF, suggesting that IR systems that utilise graded relevance assessments should be evaluated using IR metrics that utilise graded relevance, as MAP can never appreciate a system that ranks highly relevant documents above partially relevant ones.

References

[1] Amati, G. and Crestani, F.: Probabilistic Learning for Selective Dissemination of Information, *Information Processing and Management*, 35, pp. 633-654, 1999.

[2] Buckley, C. and Voorhees, E. M.: Retrieval Evaluation with Incomplete Information, *ACM SIGIR 2004 Proceedings*, pp. 25-32, 2004.

[3] Kazai, G. and Lalmas, M.: eXtended Cumulated Gain Measures for the Evaluation of Content-oriented XML Retrieval, *ACM TOIS*, 24(4), pp. 503-542, 2006.

[4] Kishida, K. *et al.*: Overview of CLIR Task at the Sixth NTCIR Workshop, *NTCIR-6 Proceedings*, 2007.

[5] Robertson, S.: On GMAP - And Other Transformations, *ACM CIKM 2006 Proceedings*, 2006.

[6] Sakai, T.: Advanced Technologies for Information Access, *International Journal of Computer Processing of Oriental Languages*, 18(2), pp. 95-113, 2005.

[7] Sakai, T., Manabe, T. and Koyama, M: Flexible Pseudo-Relevance Feedback via Selective Sampling, *ACM Transactions on Asian Language Information Processing*, 4(2), pp. 111-135, 2005.

[8] Sakai, T. *et al.*: Toshiba BRIDJE at NTCIR-5: Evaluation using Geometric Means, *NTCIR-5 Proceedings*, pp. 56-63, 2005.

[9] Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *ACM SIGIR 2006 Proceedings*, pp. 525-532, 2006.

[10] Sakai, T.: A Note on Progress in Document Retrieval Technology based on the Official NTCIR Results (in Japanese), *Forum on Information Technology 2006 Information Technology Letters*, pp. 67-70, 2006.

[11] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, 43(2), pp. 531-548, 2007.

[12] Sakai, T.: On the Reliability of Factoid Question Answering Evaluation, *ACM Transactions on Asian Language Information Processing*, 6(1), 2007.

[13] Sakai, T.: On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, *EVIA 2007* (NTCIR-6 Pre-Meeting Workshop), to appear, 2007.

[14] Sakai, T.: Alternatives to Bpref, *ACM SIGIR Proceedings*, to appear, 2007.

[15] Sparck Jones, K., Walker, S. and Robertson, S. E.: A Probabilistic Model of Information Retrieval: Development and Comparative Experiments, *Information Processing and Management* 36, Part I (pp. 779-808) and Part II (pp. 809-840), 2000.

[16] Sumner, R. G. *et al.*: Interactive Retrieval using IRIS: TREC-6 Experiments, *TREC-6 Proceedings*, 1998.

[17] Yilmaz, E. and Aslam, J. A.: Estimating Average Precision with Incomplete and Imperfect Judgments, *CIKM 2006 Proceedings*, 2006.