Sampling Precision to Depth 9000: Evaluation Experiments at NTCIR-6

Stephen Tomlinson
Open Text Corporation
Ottawa, Ontario, Canada
stomlins@opentext.com

March 16, 2007

Abstract

We describe evaluation experiments conducted by submitting retrieval runs for the Chinese, Japanese and Korean Single Language Information Retrieval subtasks of the Cross-Lingual Information Retrieval (CLIR) Task of the 6th NII Test Collection for IR Systems Workshop (NTCIR-6). We show that a Generalized Success@10 measure exposes a downside of the blind feedback technique that is overlooked by traditional ad hoc retrieval measures such as mean average precision, R-precision and Precision@10. Hence an important retrieval scenario, seeking just one item to answer a question, is not properly evaluated by the traditional ad hoc retrieval measures. Also, for each language, we submitted a one-percent subset of the first 9000 retrieved items to investigate the frequency of relevant items at deeper ranks than the official judging depth of 100. The results suggest that, on average, less than 60% of the relevant items for Chinese and less than 80% for Japanese are assessed. **Keywords:** *Chinese (Traditional), Japanese, Korean, evaluation, robust retrieval, sampling.*

1 Introduction

Livelink ECM - eDOCS SearchServer™ (formerly known as Hummingbird SearchServer™) is a toolkit for developing enterprise search and retrieval applications. The SearchServer kernel is also embedded in other components of the Livelink ECM - eDOCS Suite[1].

SearchServer works in Unicode internally [3] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (NTCIR [5], CLEF [2] and TREC [8]) have provided judged test collections for objective experimentation with SearchServer in more than a dozen languages.

This paper describes experimental work with SearchServer for the monolingual CJK tasks of finding relevant documents for natural language queries in Chinese, Japanese and Korean using the NTCIR-6 test collections.

2 Methodology

2.1 Data

The document sets of the NTCIR-6 test collections (CLIR task) were the same as for NTCIR-5. They consisted of news articles from 2000 and 2001 in Chinese (Traditional), Japanese and Korean. Table 1 gives their sizes. For more details, see the CLIR task overview paper [4].

Table 1. Sizes of NTCIR-6 Document Sets

Language	Text Size	#Documents
Chinese	1,113,487,231 bytes	901,446
Japanese	1,078,183,238 bytes	858,400
Korean	333,320,195 bytes	220,374

The NTCIR organizers re-used 50 natural language "topics" from NTCIR-3 and NTCIR-4 but, as the document sets were different, produced a new set of relevance assessments (qrels). The qrels list the documents judged to be highly relevant, relevant, partially relevant or not relevant for each of the topics. In this paper, except where otherwise stated, we just count 'highly relevant' and 'relevant' as relevant. Table 2 gives the final number of topics for each language and their average number of relevant documents (along with the lowest, median and highest number of relevant documents of the topics).

[1] Livelink, Open Text™ and SearchServer™ are trademarks or registered trademarks of Open Text Corporation in the United States of America, Canada, the European Union and/or other countries. This list of trademarks is not exhaustive. Other trademarks, registered trademarks, product names, company names, brands and service names mentioned herein are property of Open Text Corporation or other respective owners.

Table 2. Judged Topics of NTCIR-6

Language	Topics	Rel/Topic (H+R)
Chinese	50	52 (lo 8, med 41.5, hi 226)
Japanese	50	64 (lo 4, med 43, hi 210)
Korean	50	46 (lo 6, med 24.5, hi 186)

2.2 Indexing

We used word-based indexing for each language (no n-gram runs this year). For Chinese and Japanese, SearchServer segmented the text into words and split compound words (decompounding). The segmenter also performed stemming for Japanese. For Korean, SearchServer decompounded and stemmed the Korean words. A short stopword list was used for each language. The lexicon-based segmenters and stemmers were based on internal linguistic component 3.7.4.3.

2.3 Searching

For all runs, SearchServer Intuitive Searching was used, i.e. the IS_ABOUT predicate of SearchSQL, which accepts unstructured text. For example, if the Title for a topic was “地震, 台湾” (Earthquakes, Taiwan), then a corresponding SearchSQL query would be:

```
SELECT RELEVANCE() AS REL, DOCNO
FROM NTC6J
WHERE FT_TEXT IS_ABOUT '地震, 台湾'
ORDER BY REL DESC;
```

The relevance calculation included term frequency dampening [6] and inverse document frequency.

For the blind feedback runs investigated below, 3 additional IS_ABOUT queries were issued (one for each of first 3 documents retrieved by the base run). Then the 3 result lists were merged with the base result list based on the relevance scores (weight 1 for each expansion query, weight 3 for the base run).

2.4 Evaluation Measures

This paper refers to the following retrieval measures:

Precision@n: For a topic, “precision” is the percentage of retrieved documents which are relevant. “Precision@n” is the precision after n documents have been retrieved. This paper looks at Precision@10 (P10) for most runs.

R-Precision (R-Prec): For a topic, R-Prec is the precision at rank R, where R is the number of relevant items for the topic.

Average Precision (AP): For a topic, AP is the average of the precision after each relevant item is retrieved (using zero as the precision for relevant items which are not retrieved). In this paper, AP is based on the first 1000 retrieved items. The score ranges from 0.0 (no relevants found) to 1.0 (all relevants found at the top of the list). “Mean Average Precision” (MAP) is the mean of the average precision scores over all of the topics (i.e. all topics are weighted equally).

Geometric MAP (GMAP): GMAP (introduced in [17]) is based on “Log Average Precision” which for a topic is the natural log of the max of 0.00001 and the average precision. GMAP is the exponential of the mean log average precision.

Reciprocal Rank (RR): For a topic, RR is $\frac{1}{r}$ where r is the rank of the first relevant item, or zero if no relevant item is retrieved. “Mean Reciprocal Rank” (MRR) is the mean of the reciprocal ranks over all the topics.

Success@n (S@n): For a topic, Success@n is 1 if the first relevant item is found in the first n rows, 0 otherwise. This paper looks at Success@1 (S1) and Success@10 (S10) for most runs.

Generalized Success@10 (GenS@10, GenS10 or GS10): For a topic, GenS10 is 1.08^{1-r} where r is the rank of the first relevant item, or zero if no relevant item is retrieved. (This measure was introduced in [14] as “First Relevant Score” (FRS).) Compared to reciprocal rank, GenS10 falls less sharply in the early ranks; e.g. GenS10 is 1.0 at rank 1, 0.93 at rank 2, 0.86 at rank 3, etc. GenS10 is considered a generalization of Success@10 because it rounds to 1 for $r \leq 10$ and to 0 for $r > 10$.

In [12], we showed that the traditional precision measures are not satisfactory for the scenario of seeking just one item (such as to answer a question). e.g. Precision@10 favored the wrong approaches because of its weight on secondary recall, and Reciprocal Rank was erratic from overemphasizing some small differences in rank. But the GenS10 measure reflects this scenario well. Unlike for RR, a large difference in GenS10 implies an important large difference in the rank of the first relevant item. In particular, for a topic, a difference in GenS10 exceeding 0.5 implies a difference of rank of at least 10 and that one system retrieved the result in the first 10 ranks and the other did not. For sets of topics, [12] found that mean GenS10 was the most reliable of 30 investigated retrieval measures at favoring the more robust system. (In Section 3, we confirm that this result also holds for the above measures on the NTCIR data.)

2.5 Comparision Tables

For comparison tables such as Tables 4 and 5, the columns are as follows:

“Expt” specifies the language code and topic field

Table 3. Mean Scores of Stage 1 Runs

Run	GenS10	S1	MAP
HUM-C-C-D-03	0.769	17/50	0.186
HUM-C-C-D-05	0.743	18/50	0.227
HUM-C-C-T-02	0.796	18/50	0.200
HUM-C-C-T-04	0.758	15/50	0.225
HUM-J-J-D-03	0.795	22/50	0.215
HUM-J-J-D-05	0.747	19/50	0.226
HUM-J-J-T-02	0.844	22/50	0.239
HUM-J-J-T-04	0.794	25/50	0.260
HUM-K-K-D-03	0.920	32/50	0.290
HUM-K-K-D-05	0.899	32/50	0.332
HUM-K-K-T-02	0.870	30/50	0.325
HUM-K-K-T-04	0.851	30/50	0.363

Table 4. Mean Impact of Blind Feedback

Expt	ΔGS10	(95% Conf)	vs.
C-D	−0.026	(−0.06, 0.01)	10-21-19
C-T	−0.038	(−0.08, 0.00)	11-19-20
J-D	−0.048	(−0.10, 0.01)	13-22-15
J-T	−0.050	(−0.11, 0.01)	13-12-25
K-D	−0.020	(−0.05, 0.01)	8-13-29
K-T	−0.019	(−0.05, 0.01)	7-10-33
	ΔMAP		
C-D	0.042	(0.02, 0.07)	38-12-0
C-T	0.025	(0.00, 0.05)	32-18-0
J-D	0.011	(−0.01, 0.03)	27-22-1
J-T	0.020	(0.00, 0.04)	31-19-0
K-D	0.042	(0.01, 0.08)	34-16-0
K-T	0.038	(0.01, 0.06)	31-18-1

of the experiment.

"Δ" is the blind feedback score minus the base score for the specified measure.

"95% Conf" is an approximate 95% confidence interval for the mean difference (calculated from plus/minus twice the standard error of the mean difference). If zero is not in the interval, the result is "statistically significant" (at the 5% level).

"vs." is the number of topics on which the blind feedback score was higher, lower and tied (respectively) than the score of the base run. These numbers should always add to the number of topics (50).

"3 Extreme Diffs (Topic)" lists 3 of the individual topic differences, each followed by the topic number in brackets. The first difference is the largest one of any topic (based on the absolute value). The third difference is the largest difference in the other direction (so the first and third differences give the *range* of differences observed in this experiment). The middle difference is the largest of the remaining differences (based on the absolute value).

3 Robust Retrieval Experiments

The "blind feedback" technique (also known as pseudo-relevance feedback or (more ambiguously) as query expansion) is known to be bad for robustness because of its tendency to "not help (and frequently hurt) the worst performing topics" [16].

Tables 3 and 6 list the mean scores of the following 4 runs for each language, which were submitted to NII for assessment (Stage 1) in July 2006[2]:

D-03: base Description-only run

D-05: blind feedback run (50% based on D-03, 50% based on first 3 rows retrieved by D-03)

Table 5. Per-Topic Impact of Blind Feedback

Expt	3 Extreme GS10 Diffs (Topic)
C-D	0.26 (47), 0.25 (39), −0.25 (50)
C-T	−0.40 (19), −0.29 (37), 0.21 (65)
J-D	−0.79 (100), −0.53 (16), 0.31 (18)
J-T	−0.73 (37), −0.58 (65), 0.23 (110)
K-D	−0.44 (79), −0.30 (83), 0.21 (41)
K-T	−0.55 (79), −0.37 (19), 0.07 (74)
	3 Extreme AP Diffs (Topic)
C-D	0.20 (95), 0.20 (14), −0.08 (24)
C-T	0.31 (53), 0.14 (59), −0.11 (80)
J-D	0.16 (21), 0.14 (59), −0.13 (36)
J-T	0.26 (14), 0.21 (74), −0.09 (42)
K-D	0.35 (20), 0.35 (14), −0.15 (16)
K-T	0.26 (96), 0.22 (43), −0.08 (102)

Table 6. More Mean Scores of Stage 1

Run	S10	P10	R-Prec	GMAP
C-D-03	40/50	0.258	0.223	0.122
C-D-05	40/50	0.320	0.264	0.146
C-T-02	45/50	0.304	0.245	0.138
C-T-04	43/50	0.318	0.262	0.135
J-D-03	42/50	0.340	0.243	0.093
J-D-05	39/50	0.342	0.249	0.091
J-T-02	46/50	0.376	0.263	0.140
J-T-04	39/50	0.372	0.279	0.148
K-D-03	49/50	0.408	0.300	0.215
K-D-05	47/50	0.444	0.340	0.239
K-T-02	45/50	0.434	0.329	0.176
K-T-04	44/50	0.478	0.355	0.189

[2] Hummingbird was subsequently acquired by Open Text Corporation on October 2, 2006.

Table 7. Statistically Significant Mean Differences in 6 Blind Feedback Experiments of Stage 1

Measure	Signif. Negative	Signif. Positive
GenS10	1/6	0/0
S10	1/5	0/0
MRR	0/5	0/1
S1	0/2	0/2
GMAP	0/2	1/4
P10	0/1	3/5
R-Prec	0/0	3/6
MAP	0/0	5/6

T-02: base Title-only run

T-04: blind feedback run (50% based on T-02, 50% based on first 3 rows retrieved by T-02)

Hence these runs provide 6 official blind feedback experiments (compare D-03 to D-05 and T-02 to T-04 for each of the 3 languages).

Table 3 shows that mean GenS10 *declined* in all 6 blind feedback experiments, i.e. blind feedback pushed down the first relevant item (on average). Blind feedback is not a good technique to use if you are just seeking one item to answer a question.

Table 3 shows that MAP *increased* in all 6 blind feedback experiments, i.e. MAP favors blind feedback, a non-robust technique.

Table 4 shows that for Chinese Titles (C-T), both the decline in GenS10 and increase in MAP were statistically significant.

Table 5 shows that the range of per-topic differences was larger for GenS10 than AP. Blind feedback can have a large, negative impact on the first relevant item for individual topics.

Table 6 shows that in 4 of the cases in which P10 was up with blind feedback, S10 still fell. (P10 and S10 can be very different.)

Table 7 summarizes how many of the mean differences were negative and positive for each retrieval measure, and which of these differences were statistically significant. (For example, it shows for P10 that 3 of the 5 positive mean differences were statistically significant.) The overall pattern is that measures based on just the first relevant item (such as GenS10, S10 and MRR) declined with blind feedback, while traditional ad hoc measures (such as MAP, R-Prec and P10) increased with blind feedback.

These results are consistent with what we have seen elsewhere [14, 10, 13, 12, 15]. For example, in [12], 7 other groups' blind feedback systems (of the 2003 RIA workshop) were studied, and it was found that blind feedback was detrimental to the first relevant item (on average), even though it boosted the traditional measures. (So this result is not particular to how we do blind feedback.)

[1] gives a theoretical explanation for why different retrieval approaches are superior when seeking just one relevant item instead of several. In particular, it finds that when seeking just one relevant item, it can theoretically be advantageous to use *negative* pseudo-relevance feedback to encourage more diversity in the results (i.e. after retrieving the first item, assume it is *not* relevant when deciding what to retrieve next; duplicate filtering is a special case of negative feedback).

Intuitively, the reason that the traditional measures (such as P10, R-Prec and MAP) are not robust is that they encourage retrieval of duplicate information (and penalize duplicate filtering).

3.1 Stage 2 Results

For Stage 2, the organizers requested that we participants check the consistency of our results by submitting runs with our latest systems using the test topics and document sets of each of the previous 3 rounds of NTCIR (labelled N3, N4 and N5). The old qrels of these forums were re-used in Stage 2 (no new assessing was done).

We submitted the following 5 runs for each language in September 2006:

D-02: same approach as our Stage 1 D-03 run (base D run)

D-05: same approach as our Stage 1 D-05 run (blind feedback run)

T-01: same approach as our Stage 1 T-02 run (base T run)

T-04: same approach as our Stage 1 T-04 run (blind feedback run)

TDNC-03: plain full topic run

These runs give us 18 more blind feedback experiments (compare D-02 to D-05 and T-01 to T-04 for each of the 3 languages, for each of the 3 rounds).

Table 8 lists the mean scores of the Stage 2 runs, and Table 9 summarizes how many of the mean differences were negative and positive for each retrieval measure, and which of these differences were statistically significant. For example, it shows that 3 of the 14 negative differences for GenS10 were statistically significant. Overall, the patterns are the same as for Stage 1.

[7] recently made the (unsupported) claim that for GMAP, "blind feedback is often found to be detrimental". However, in our past official experiments with GMAP ([13, 11]) and in the RIA experiments ([12]) we have seen statistically significant increases in GMAP from blind feedback, but no statistically significant decreases. We see this result again in Table 9; all 5 statistically significant differences for GMAP favored blind feedback. We do not consider GMAP to be a robust measure.

Table 8. Mean Scores of Stage 2 Runs

Run	GenS10	S1	MAP
HUM-C-C-D-02-N3	0.683	13/42	0.190
HUM-C-C-D-05-N3	0.615	9/42	0.199
HUM-C-C-T-01-N3	0.643	11/42	0.210
HUM-C-C-T-04-N3	0.649	16/42	0.238
HUM-J-J-D-02-N3	0.812	21/42	0.307
HUM-J-J-D-05-N3	0.800	21/42	0.320
HUM-J-J-T-01-N3	0.802	24/42	0.327
HUM-J-J-T-04-N3	0.804	21/42	0.346
HUM-K-K-D-02-N3	0.744	15/30	0.244
HUM-K-K-D-05-N3	0.740	13/30	0.250
HUM-K-K-T-01-N3	0.745	15/30	0.263
HUM-K-K-T-04-N3	0.696	13/30	0.277
-C-C-TDNC-03-N3	0.786	16/42	0.268
-J-J-TDNC-03-N3	0.848	20/42	0.371
-K-K-TDNC-03-N3	0.825	17/30	0.326
HUM-C-C-D-02-N4	0.683	11/59	0.154
HUM-C-C-D-05-N4	0.651	16/59	0.172
HUM-C-C-T-01-N4	0.719	17/59	0.169
HUM-C-C-T-04-N4	0.719	18/59	0.177
HUM-J-J-D-02-N4	0.911	25/55	0.300
HUM-J-J-D-05-N4	0.923	34/55	0.321
HUM-J-J-T-01-N4	0.930	32/55	0.307
HUM-J-J-T-04-N4	0.906	35/55	0.339
HUM-K-K-D-02-N4	0.875	35/57	0.347
HUM-K-K-D-05-N4	0.812	30/57	0.390
HUM-K-K-T-01-N4	0.928	38/57	0.379
HUM-K-K-T-04-N4	0.897	34/57	0.423
-C-C-TDNC-03-N4	0.776	24/59	0.205
-J-J-TDNC-03-N4	0.937	36/55	0.340
-K-K-TDNC-03-N4	0.923	34/57	0.407
HUM-C-C-D-02-N5	0.820	25/50	0.270
HUM-C-C-D-05-N5	0.801	20/50	0.327
HUM-C-C-T-01-N5	0.871	31/50	0.324
HUM-C-C-T-04-N5	0.839	29/50	0.355
HUM-J-J-D-02-N5	0.815	18/47	0.282
HUM-J-J-D-05-N5	0.788	19/47	0.290
HUM-J-J-T-01-N5	0.885	28/47	0.312
HUM-J-J-T-04-N5	0.863	23/47	0.338
HUM-K-K-D-02-N5	0.899	34/50	0.354
HUM-K-K-D-05-N5	0.886	33/50	0.416
HUM-K-K-T-01-N5	0.912	29/50	0.355
HUM-K-K-T-04-N5	0.905	32/50	0.425
-C-C-TDNC-03-N5	0.875	29/50	0.372
-J-J-TDNC-03-N5	0.890	27/47	0.375
-K-K-TDNC-03-N5	0.946	38/50	0.455

Table 9. Statistically Significant Mean Differences in 18 Blind Feedback Experiments of Stage 2

Measure	Signif. Negative	Signif. Positive
GenS10	3/14	0/4
S10	0/12	0/1
MRR	3/11	1/7
S1	2/10	2/7
GMAP	0/3	5/15
P10	0/2	7/16
R-Prec	0/2	6/16
MAP	0/0	10/18

Note that for experiments not involving feedback techniques, GenS10 and MAP tend to agree. e.g. The TDNC score is usually higher than the T or D score for both GenS10 and MAP.

4 Precision to Depth 9000

For Stage 1, our submitted full topic (TDNC) run for each language was actually a depth probe run from sampling a plain base-TDNC run for the language (the base-TDNC run was not itself submitted because of the 5-run submission limit). The submitted TDNC run contained rows 1, 101, 201, 301, ..., 9001 of the base-TDNC run, followed by rows 9002, 9003, 9004, ..., 9910 of the base-TDNC run. This run was given highest judging precedence, and the first 100 rows were judged, allowing us to investigate precision rates to depth 9000 for each language. (It appears actually that all submitted runs were judged to depth 100.)

Tables 10, 12 and 14 list the number of each type of item retrieved for each depth range, for Chinese, Japanese and Korean respectively, over the 50 topics. The item type codes are H (highly relevant), R (ordinary relevant), P (partially relevant), N (non-relevant) and U (unjudged). All depth ranges contain 10 sample points (except the last one which just contains 9); hence over 50 topics, there are 500 items summarized in each row (450 for the last row). Tables 11, 13 and 15 list the marginal precision rates for each depth range, both for rigid relevance (H+R) and relaxed relevance (H+R+P).

For example, Table 10 shows that for the 10 depth points from 101-1001, 9 highly relevant items and 15 ordinary relevant items were retrieved in the 500 items, hence the marginal precision rate was 4.8% ((9+15)/500) over this range, which is listed in the corresponding row of Table 11. This precision rate suggests that there are dozens of relevant items (per topic) deeper than the usual judging depth of 100. (However, after depth 2000, marginal precision was less than 1%

Table 10. Relevant Items of Chinese Base-TDNC Run at Various Depths

Depths	#Rel (over 50 Topics)
1, 2, ..., 10	100H, 75R, 85P, 236N, 4U
11, 12, ..., 20	64H, 69R, 55P, 309N, 3U
21, 22, ..., 30	63H, 53R, 62P, 310N, 12U
31, 32, ..., 40	50H, 58R, 57P, 317N, 18U
41, 42, ..., 50	35H, 43R, 54P, 340N, 28U
51, 52, ..., 60	30H, 45R, 38P, 357N, 30U
61, 62, ..., 70	41H, 38R, 52P, 344N, 25U
71, 72, ..., 80	31H, 39R, 38P, 351N, 41U
81, 82, ..., 90	20H, 40R, 26P, 362N, 52U
91, 92, ..., 100	26H, 39R, 44P, 337N, 54U
101, 201, ..., 1001	9H, 15R, 10P, 466N, 0U
1101, 1201, ..., 2001	2H, 5R, 1P, 492N, 0U
2101, 2201, ..., 3001	2H, 2R, 2P, 494N, 0U
3101, 3201, ..., 4001	4H, 0R, 5P, 491N, 0U
4101, 4201, ..., 5001	1H, 1R, 0P, 498N, 0U
5101, 5201, ..., 6001	2H, 2R, 2P, 494N, 0U
6101, 6201, ..., 7001	0H, 0R, 1P, 499N, 0U
7101, 7201, ..., 8001	0H, 0R, 0P, 500N, 0U
8101, 8201, ..., 9001	0H, 0R, 0P, 500N, 0U
9002, 9003, ..., 9010	0H, 0R, 0P, 450N, 0U

for all 3 languages.)

Based on the 90 sample points from depths 101-9001, we can form an estimate of the number of relevant items retrieved in the first 9000 rows for each language. (In fact, the estimate is probably low, because we are in effect using depth 101 to represent rows 1-100, depth 201 to represent rows 101-200, etc.) For example, Table 10 shows that the sample for Chinese includes 20 highly relevant items (9+2+2+4+1+2+0+0+0) and 25 ordinary relevant items (15+5+2+0+1+2+0+0+0), or 45 relevant items (20+25), which projects to 4500 relevant items retrieved in the first 9000 rows (45*(9000/90)) over the 50 topics, or 90 relevant items per topic (4500/50). Hence 90 is the listed “Estimated H+R@9000” result for Chinese in Table 16. Table 16 also lists the corresponding numbers for Japanese and Korean.

The “Official H+R” row of Table 16 lists the official number of relevant items per topic for each language. For example, for Chinese, the official qrels contain 52 relevant items on average per topic.

The “Percentage Judged” row of Table 16 just divides the official number of relevant items by the estimated number in the first 9000 retrieved (e.g. for Chinese, 52/90=58%). This number is likely an overestimate of the percentage of all relevant items that are judged (on average per topic) because there may be relevant items that were not matched by the query in the first 9000 rows.

Table 11. Marginal Precision of Chinese Base-TDNC Run at Various Depths

Depths	Prec H+R	Prec H+R+P
1, 2, ..., 10	0.350	0.520
11, 12, ..., 20	0.266	0.376
21, 22, ..., 30	0.232	0.356
31, 32, ..., 40	0.216	0.330
41, 42, ..., 50	0.156	0.264
51, 52, ..., 60	0.150	0.226
61, 62, ..., 70	0.158	0.262
71, 72, ..., 80	0.140	0.216
81, 82, ..., 90	0.120	0.172
91, 92, ..., 100	0.130	0.218
101, 201, ..., 1001	0.048	0.068
1101, 1201, ..., 2001	0.014	0.016
2101, 2201, ..., 3001	0.008	0.012
3101, 3201, ..., 4001	0.008	0.018
4101, 4201, ..., 5001	0.004	0.004
5101, 5201, ..., 6001	0.008	0.012
6101, 6201, ..., 7001	0.000	0.002
7101, 7201, ..., 8001	0.000	0.000
8101, 8201, ..., 9001	0.000	0.000
9002, 9003, ..., 9010	0.000	0.000

Table 12. Relevant Items of Japanese Base-TDNC Run at Various Depths

Depths	#Rel (over 50 Topics)
1, 2, ..., 10	23H, 174R, 67P, 235N, 1U
11, 12, ..., 20	5H, 130R, 63P, 299N, 3U
21, 22, ..., 30	6H, 110R, 54P, 320N, 10U
31, 32, ..., 40	4H, 98R, 53P, 337N, 8U
41, 42, ..., 50	5H, 88R, 43P, 346N, 18U
51, 52, ..., 60	5H, 102R, 39P, 332N, 22U
61, 62, ..., 70	7H, 87R, 39P, 337N, 30U
71, 72, ..., 80	6H, 81R, 42P, 347N, 24U
81, 82, ..., 90	3H, 73R, 32P, 353N, 39U
91, 92, ..., 100	5H, 76R, 30P, 335N, 54U
101, 201, ..., 1001	0H, 27R, 17P, 456N, 0U
1101, 1201, ..., 2001	0H, 7R, 3P, 490N, 0U
2101, 2201, ..., 3001	0H, 4R, 2P, 494N, 0U
3101, 3201, ..., 4001	0H, 0R, 0P, 500N, 0U
4101, 4201, ..., 5001	0H, 0R, 0P, 500N, 0U
5101, 5201, ..., 6001	0H, 0R, 1P, 499N, 0U
6101, 6201, ..., 7001	0H, 0R, 1P, 499N, 0U
7101, 7201, ..., 8001	0H, 3R, 0P, 497N, 0U
8101, 8201, ..., 9001	0H, 0R, 0P, 500N, 0U
9002, 9003, ..., 9010	0H, 0R, 0P, 450N, 0U

Table 13. Marginal Precision of Japanese Base-TDNC Run at Various Depths

Depths	Prec H+R	Prec H+R+P
1, 2, ..., 10	0.394	0.528
11, 12, ..., 20	0.270	0.396
21, 22, ..., 30	0.232	0.340
31, 32, ..., 40	0.204	0.310
41, 42, ..., 50	0.186	0.272
51, 52, ..., 60	0.214	0.292
61, 62, ..., 70	0.188	0.266
71, 72, ..., 80	0.174	0.258
81, 82, ..., 90	0.152	0.216
91, 92, ..., 100	0.162	0.222
101, 201, ..., 1001	0.054	0.088
1101, 1201, ..., 2001	0.014	0.020
2101, 2201, ..., 3001	0.008	0.012
3101, 3201, ..., 4001	0.000	0.000
4101, 4201, ..., 5001	0.000	0.000
5101, 5201, ..., 6001	0.000	0.002
6101, 6201, ..., 7001	0.000	0.002
7101, 7201, ..., 8001	0.006	0.006
8101, 8201, ..., 9001	0.000	0.000
9002, 9003, ..., 9010	0.000	0.000

Table 14. Relevant Items of Korean Base-TDNC Run at Various Depths

Depths	#Rel (over 50 Topics)
1, 2, ..., 10	151H, 84R, 112P, 150N, 3U
11, 12, ..., 20	93H, 62R, 98P, 242N, 5U
21, 22, ..., 30	72H, 57R, 90P, 267N, 14U
31, 32, ..., 40	64H, 47R, 94P, 270N, 25U
41, 42, ..., 50	57H, 31R, 79P, 311N, 22U
51, 52, ..., 60	46H, 29R, 82P, 310N, 33U
61, 62, ..., 70	49H, 22R, 70P, 323N, 36U
71, 72, ..., 80	42H, 16R, 61P, 332N, 49U
81, 82, ..., 90	45H, 15R, 53P, 326N, 61U
91, 92, ..., 100	45H, 23R, 54P, 295N, 83U
101, 201, ..., 1001	10H, 4R, 17P, 469N, 0U
1101, 1201, ..., 2001	4H, 1R, 3P, 492N, 0U
2101, 2201, ..., 3001	0H, 0R, 2P, 498N, 0U
3101, 3201, ..., 4001	0H, 0R, 2P, 498N, 0U
4101, 4201, ..., 5001	0H, 0R, 1P, 499N, 0U
5101, 5201, ..., 6001	1H, 0R, 3P, 496N, 0U
6101, 6201, ..., 7001	1H, 0R, 1P, 498N, 0U
7101, 7201, ..., 8001	1H, 0R, 0P, 499N, 0U
8101, 8201, ..., 9001	0H, 1R, 1P, 498N, 0U
9002, 9003, ..., 9010	0H, 0R, 3P, 447N, 0U

Table 15. Marginal Precision of Korean Base-TDNC Run at Various Depths

Depths	Prec H+R	Prec H+R+P
1, 2, ..., 10	0.470	0.694
11, 12, ..., 20	0.310	0.506
21, 22, ..., 30	0.258	0.438
31, 32, ..., 40	0.222	0.410
41, 42, ..., 50	0.176	0.334
51, 52, ..., 60	0.150	0.314
61, 62, ..., 70	0.142	0.282
71, 72, ..., 80	0.116	0.238
81, 82, ..., 90	0.120	0.226
91, 92, ..., 100	0.136	0.244
101, 201, ..., 1001	0.028	0.062
1101, 1201, ..., 2001	0.010	0.016
2101, 2201, ..., 3001	0.000	0.004
3101, 3201, ..., 4001	0.000	0.004
4101, 4201, ..., 5001	0.000	0.002
5101, 5201, ..., 6001	0.002	0.008
6101, 6201, ..., 7001	0.002	0.004
7101, 7201, ..., 8001	0.002	0.002
8101, 8201, ..., 9001	0.002	0.004
9002, 9003, ..., 9010	0.000	0.007

Table 16. Estimated Percentage of Relevant Items that are Judged, Per Topic

	C	J	K
Estimated H+R@9000	90	82	46
Official H+R	52	64	46
Percentage Judged	58%	78%	100%
Estimated H+R+P@9000	132	130	106
Official H+R+P	88	95	90
Percentage Judged	67%	73%	85%

The estimated judging coverage for the NTCIR-6 collections (58% for Chinese, 78% for Japanese, 100% for Korean) is much higher than the estimates we produced for the TREC 2006 Legal and Terabyte collections using a similar approach (18% for TREC Legal and 36% for TREC Terabyte) [15].

There is a lot of variance in the estimates across the topics. For example, for Korean topic 37, 5 of our 90 sample points were judged relevant, which projects to 500 relevant items in the first 9000 retrieved, whereas the offical qrels contained 94 relevant items for this topic, suggesting that at least 406 relevant items are not judged. (Hence probably less than 100% of Korean relevant items are actually judged. For some topics, our sample is obviously missing a lot of relevant items; e.g. for Korean topic 105, 0 of our 90 sample points were judged relevant, but the official qrels contain 120 relevant items for this topic.)

A pattern seems to be that the more official relevant items there are for a topic, the more likely it is there are a lot of unjudged relevant items for the topic. For example, topic 64 has the most official relevant items for Chinese (226), and 5 of our 90 sample points were judged relevant for this topic (suggesting at least 500 relevant items for the topic). Likewise, topic 41 has the most official relevant items for Japanese (210), and 4 of our 90 sample points were judged relevant for this topic (suggesting at least 400 relevant items for the topic).

Another pattern seems to be that larger collections have less judging coverage than smaller collections. e.g. the judging coverage for Chinese and Japanese (almost 1 million documents each) is less than for Korean (approximately 200,000 documents), and the judging coverage is even less for the larger TREC collections (approximately 7 million documents for Legal, 25 million for Terabyte).

These incompleteness results are similar to what [18] found for depth-100 pooling on the old TREC collections of approximately 500,000 documents: “it is likely that at best 50%-70% of the relevant documents have been found; most of these unjudged relevant documents are for the 10 or so queries that already have the most known answers.”

Fortunately, [18] also found for such test collections that “overall they do indeed lead to reliable results.” (We can also confirm that we have gained a lot of insights from the NTCIR test collections over the years, particularly from the topic analyses in [9, 10].)

5 Conclusions

We have conducted a sampling experiment which has helped us put our recent results on large TREC collections in context. We have demonstrated that, on average per topic, less than 60% of the Chinese relevant items and less than 80% of the Japanese relevant items are assessed. This judging coverage is actually much higher than we found for the (much larger) TREC 2006 Legal collection (less than 20% of relevant items assessed). We also produced tables which showed the frequency of relevant items down to depth 9000 (well beyond the official judging depth of 100). It appears that dozens of relevant items can still be retrieved beyond depth 100, on average, though marginal precision of our full topic run was below 1% after depth 2000 for all 3 languages. (In contrast, marginal precision was still above 4% at depth 9000 in our sampled TREC Legal run).

We have also confirmed on the NTCIR collections that measures based on the rank of the first relevant item, particularly the Generalized Success@10 measure, expose a downside of the blind feedback technique that is not reflected in the traditional ad hoc retrieval measures (such as mean average precision, R-precision and Precision@10). Hence an important retrieval scenario, seeking just one item to answer a question, is not properly evaluated by the traditional ad hoc retrieval measures. We believe that more attention should be given to measures of the first relevant item in ad hoc evaluations, particularly to the GenS@10 measure.

References

[1] H. Chen and D. R. Karger. Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. *Proceedings of SIGIR 2006*, pp. 429-436.

[2] Cross-Language Evaluation Forum (CLEF) web site. http://www.clef-campaign.org/.

[3] A. Hodgson. Converting the Fulcrum Search Engine to Unicode. *Sixteenth International Unicode Conference*, 2000.

[4] K. Kishida, Kuang-hua Chen, S. Lee, K. Kuriyama, N. Kando, Hsin-Hsi Chen and S. H. Myaeng. Overview of CLIR Task at the Sixth NTCIR Workshop. To appear in *Proceedings of NTCIR-6*, 2007.

[5] NTCIR (NII Test Collection for IR Systems) Project Home Page. http://research.nii.ac.jp/ntcir/.

[6] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu and M. Gatford. Okapi at TREC-3. *Proceedings of TREC-3*, 1995.

[7] S. Robertson. On GMAP – and other transformations. *Proceedings of CIKM 2006*, pp. 78-83.

[8] Text REtrieval Conference (TREC) Home Page. http://trec.nist.gov/.

[9] S. Tomlinson. CJK Experiments with Hummingbird SearchServer^{TM} at NTCIR-4. *Proceedings of NTCIR-4*, 2004.

[10] S. Tomlinson. CJK Experiments with Hummingbird SearchServer^{TM} at NTCIR-5. *Proceedings of NTCIR-5*, 2005.

[11] S. Tomlinson. Comparing the Robustness of Expansion Techniques and Retrieval Measures. *Working Notes for the CLEF 2006 Workshop*.

[12] S. Tomlinson. Early Precision Measures: Implications from the Downside of Blind Feedback. *Proceedings of SIGIR 2006*, pp. 705-706.

[13] S. Tomlinson. Enterprise, QA, Robust and Terabyte Experiments with Hummingbird SearchServer^{TM} at TREC 2005. *Proceedings of TREC 2005*.

[14] S. Tomlinson. European Ad Hoc Retrieval Experiments with Hummingbird SearchServer^{TM} at CLEF 2005. *Working Notes for the CLEF 2005 Workshop*.

[15] S. Tomlinson. Experiments with the Negotiated Boolean Queries of the TREC 2006 Legal Discovery Track. *Proceedings of TREC 2006*.

[16] E. M. Voorhees. Overview of the TREC 2003 Robust Retrieval Track. *Proceedings of TREC 2003*.

[17] E. M. Voorhees. Overview of the TREC 2004 Robust Retrieval Track. *Proceedings of TREC 2004*.

[18] J. Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments? *Proceedings of SIGIR’98*, pp. 307-314.