Sampling Precision to Depth 10000: Evaluation Experiments at CLEF 2007

Stephen Tomlinson
Open Text Corporation
Ottawa, Ontario, Canada
stomlins@opentext.com
http://www.opentext.com/

August 20, 2007

Abstract

We describe evaluation experiments conducted by submitting retrieval runs for the monolingual Bulgarian, Czech and Hungarian information retrieval tasks of the Ad-Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2007. In the ad hoc retrieval tasks, the system was given 50 natural language queries, and the goal was to find all of the relevant documents (with high precision) in a particular document set. We conducted diagnostic experiments with different techniques for matching word variations and handling stopwords, comparing the performance on the robust Generalized Success@10 measure and the non-robust mean average precision measure. The measures generally agreed on the mean benefits of morphological techniques such as stemming, but generally disagreed on the blind feedback technique, though not all of the mean differences were statistically significant. Also, for each language, we submitted a sample of the first 10000 retrieved items to investigate the frequency of relevant items at deeper ranks than the official judging depth (of 60 for Czech and 80 for Bulgarian and Hungarian). The results suggest that, on average, the percentage of relevant items assessed was less than 60% for Czech, 70% for Bulgarian and 85% for Hungarian.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Bulgarian Retrieval, Czech Retrieval, Hungarian Retrieval, Robust Retrieval, Sampling

1 Introduction

Livelink ECM - eDOCS SearchServer^{TM} (formerly known as Hummingbird SearchServer^{TM}) 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].

[1] Livelink, Open Text^{TM} and SearchServer^{TM} 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 1: Sizes of CLEF 2007 Ad-Hoc Track Test Collections

Language	Text Size (uncompressed)	Documents	Topics	Rel/Topic
Bulgarian	265,368,055 bytes	87,281	50	20 (lo 2, hi 62)
Czech	151,914,429 bytes	81,735	50	15 (lo 2, hi 47)
Hungarian	106,631,823 bytes	49,530	50	18 (lo 1, hi 66)

SearchServer works in Unicode internally [4] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (CLEF [3], NTCIR [5] 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 task of finding relevant documents for natural language queries in various European languages using the CLEF 2007 Ad-Hoc Track test collections.

2 Methodology

2.1 Data

The CLEF 2007 Ad-Hoc Track document sets consisted of tagged (SGML-formatted) news articles in 3 different languages: Bulgarian, Czech and Hungarian. Table 1 gives the sizes.

The CLEF organizers created 50 natural language "topics" (numbered 401-450) and translated them into many languages. Sometimes topics are discarded for some languages because of a lack of relevant documents. Table 1 gives the final number of topics for each language and their average number of relevant documents (along with the lowest and highest number of relevant documents of any topic). For more information on the CLEF test collections, see the track overview paper.

2.2 Indexing

Our indexing approach was mostly the same as last year [9]. Accents were not indexed except for the combining breve in Bulgarian. The apostrophe was treated as a word separator for the investigated languages. The custom text reader, cTREC, was updated to maintain support for the CLEF guidelines of only indexing specifically tagged fields.

For some experiments, some stop words were excluded from indexing (e.g. words like "the", "by" and "of" in English). For these experiments, the stop word lists for Bulgarian, Czech and Hungarian were based on Savoy's lists [7].

By default, the SearchServer index supports both exact matching (after some Unicode-based normalizations, such as decompositions and conversion to upper-case) and morphological matching (e.g. inflections, derivations and compounds, depending on the linguistic component used).

2.3 Searching

We experimented with the SearchServer CONTAINS predicate. Our test application specified SearchSQL to perform a boolean-OR of the query words. For example, for Czech topic 405 whose Title was "Astma u dětství" (Childhood Asthma), a corresponding SearchSQL query would be:

```
SELECT RELEVANCE('2:4') AS REL, DOCNO
FROM CLEF07CS
WHERE FT_TEXT CONTAINS 'Astma'|'u'|'dětství'
ORDER BY REL DESC;
```

Most aspects of the SearchServer relevance value calculation are the same as described last year [9]. Briefly, SearchServer dampens the term frequency and adjusts for document length in a manner similar to Okapi [6] and dampens the inverse document frequency using an approximation of the logarithm. These calculations are based on the stems of the terms (roughly speaking) when doing morphological searching (i.e. when SET TERM_GENERATOR 'word!ftelp/inflect' was previously specified). The SearchServer RELEVANCE_METHOD setting was set to '2:4' and RELEVANCE_DLEN_IMP was set to 750 for all experiments in this paper.

2.4 Diagnostic Runs

For the diagnostic runs listed in Tables 2, the run names consist of a language code ("BG" for Bulgarian, "CS" for Czech and "HU" for Hungarian) followed by one of the following labels:

- "none": No linguistic variations from stemming were matched. Just the surface forms were searched on (after case-normalization).
- "stem": Same as "none" except that linguistic variations from stemming were matched. We thank Jacques Savoy for providing experimental algorithmic stemmers [7] for all 3 languages. For Czech, our port of the stemmer was accent-insensitive.
- "all": Same as "stem" except that a separate index was used which did not stop any words from being indexed.
- "4gram": Same as "all" except that the run used a different index which primarily consisted of the 4-grams of terms, e.g. the word 'search' would produce index terms of 'sear', 'earc' and 'arch'. No stemming was done; searching used the IS_ABOUT predicate (instead of the CONTAINS predicate) with morphological options disabled to search for the 4-grams of the query terms. For Bulgarian, we did not index the breve accent for the 4-gram runs (unlike 2 years ago [12]).
- "fuse": Fusion run based on adding together the rsv scores of the "stem" and "4gram" runs.

Note that all diagnostic runs just used the Title field of the topic.

2.5 Retrieval Measures

Traditionally, different retrieval measures have been used for "ad hoc" tasks, which seek relevant items for a topic, than for "known-item" tasks, which seek a particular known document. However, we argue that the known-item measures are not only applicable to ad hoc tasks, but that they are often preferable. For many ad hoc tasks, e.g. finding answer documents for questions, just one relevant item is needed. Also, the traditional ad hoc measures encourage retrieval of duplicate relevants, which does not correspond to user benefit.

The traditional known-item measures are very coarse, e.g. Success@10 is 1 or 0 for each topic, while reciprocal rank cannot produce a value between 1.0 and 0.5. Two years ago, we began investigating a new measure, Generalized Success@10 (GS10) (introduced as "First Relevant Score" (FRS) in [12]), which is defined below. This investigation led to the discovery that the blind feedback technique (a commonly used technique at CLEF, NTCIR and TREC, but not known to be popular in real systems) had the downside of pushing down the first relevant item (on average), as has now been verified not just for our own blind feedback approach, but for the 7 blind feedback systems of the 2003 RIA Workshop [10] and for the Neuchâtel system using French data from CLEF [1]. [2] provides a theoretical explanation for why positive feedback approaches are detrimental to the rank of the first relevant item.

2.5.1 Primary Recall Measures

"Primary recall" is retrieval of the first relevant item for a topic. Primary recall measures include the following:

- *Generalized Success@30* (GS30): For a topic, GS30 is 1.024^{1-r} where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found.
- *Generalized Success@10* (GS10): For a topic, GS10 is 1.08^{1-r} where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found.
- *Success@n* (S@n): For a topic, Success@n is 1 if a desired page is found in the first n rows, 0 otherwise. This paper lists Success@1 (S1) and Success@10 (S10) for all runs.
- *Reciprocal Rank* (RR): For a topic, RR is $\frac{1}{r}$ where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found. "Mean Reciprocal Rank" (MRR) is the mean of the reciprocal ranks over all the topics.

Interpretation of Generalized Success@n: GS30 and GS10 are estimates of the percentage of potential result list reading the system saved the user to get to the first relevant item, assuming that users are less and less likely to continue reading as they get deeper into the result list.

Comparison of GS10 and Reciprocal Rank: Both GS10 and RR are 1.0 if a desired page is found at rank 1. At rank 2, GS10 is just 7 points lower (0.93), whereas RR is 50 points lower (0.50). At rank 3, GS10 is another 7 points lower (0.86), whereas RR is 17 points lower (0.33). At rank 10, GS10 is 0.50, whereas RR is 0.10. GS10 is greater than RR for ranks 2 to 52 and lower for ranks 53 and beyond.

Connection of GS10 to Success@10: GS10 is considered a generalization of Success@10 because it rounds to 1 for $r \leq 10$ and to 0 for $r > 10$. (Similarly, GS30 is considered a generalization of Success@30 because it rounds to 1 for $r \leq 30$ and to 0 for $r > 30$.)

2.5.2 Secondary Recall Measures

"Secondary recall" is retrieval of the additional relevant items for a topic (after the first one). Secondary recall measures place most of their weight on these additional relevant items.

- *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 lists Precision@10 (P10) for all runs.
- *Average Precision* (AP): For a topic, AP is the average of the precision after each relevant document is retrieved (using zero as the precision for relevant documents which are not retrieved). By convention, AP is based on the first 1000 retrieved documents for the topic. 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 [14]) 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. (We argue in [10] that primary recall measures better reflect robustness than GMAP.)

Table 2: Mean Scores of Diagnostic Monolingual Ad Hoc Runs

Run	GS30	GS10	S10	MRR	S1	P10	GMAP	MAP
BG-4gram	0.863	0.739	38/50	0.526	20/50	0.354	0.164	0.282
BG-fuse	0.859	0.725	36/50	0.538	22/50	0.326	0.170	0.295
BG-stem	0.840	0.723	39/50	0.542	22/50	0.316	0.147	0.281
BG-all	0.837	0.716	38/50	0.521	20/50	0.312	0.146	0.282
BG-none	0.840	0.694	35/50	0.474	18/50	0.272	0.095	0.209
CS-all	0.907	0.826	44/50	0.631	25/50	0.294	0.135	0.288
CS-stem	0.905	0.823	44/50	0.633	25/50	0.294	0.135	0.289
CS-fuse	0.867	0.806	42/50	0.610	22/50	0.298	0.144	0.315
CS-4gram	0.846	0.790	42/50	0.625	25/50	0.312	0.123	0.310
CS-none	0.805	0.719	39/50	0.537	21/50	0.228	0.063	0.215
HU-fuse	0.895	0.833	44/50	0.652	26/50	0.358	0.158	0.329
HU-4gram	0.882	0.820	44/50	0.639	26/50	0.374	0.145	0.328
HU-all	0.868	0.803	42/50	0.608	23/50	0.300	0.103	0.263
HU-stem	0.866	0.801	42/50	0.608	23/50	0.300	0.103	0.263
HU-none	0.705	0.610	33/50	0.443	17/50	0.230	0.021	0.181

2.6 Statistical Significance Tables

For tables comparing 2 diagnostic runs (such as Table 3), the columns are as follows:

- "Expt" specifies the experiment. The language code is given, followed by the labels of the 2 runs being compared. The difference is the first run minus the second run. For example, "BG-stem-none" specifies the difference of subtracting the scores of the Bulgarian 'none' run from the Bulgarian 'stem' run (of Table 2).
- "ΔGS10" is the difference of the mean GS10 scores of the two runs being compared (and "ΔMAP" is the difference of the mean average precision scores).
- "95% Conf" is an approximate 95% confidence interval for the 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), i.e. the feature is unlikely to be of neutral impact (on average), though if the average difference is small (e.g. <0.020) it may still be too minor to be considered "significant" in the magnitude sense.
- "vs." is the number of topics on which the first run scored higher, lower and tied (respectively) compared to the second run. These numbers should always add to the number of topics.
- "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 Results of Morphological Experiments

3.1 Impact of Stemming

Table 3 shows the impact of stemming for the 3 languages. The mean increases in GenS@10 were statistically significant for Czech and Hungarian. For example, Table 3 shows that the biggest increase in GenS@10 for Czech was for topic 431 (Francouzští prezidenští kandidáti (French Presiden-

Table 3: Impact of Stemming on GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-stm-none	0.029	(−0.045, 0.103)	19-15-16	−0.71 (436), 0.66 (448), 0.66 (450)
CS-stm-none	0.105	(0.032, 0.177)	21-10-19	0.87 (431), 0.77 (422), −0.50 (403)
HU-stm-none	0.191	(0.088, 0.293)	22-10-18	1.00 (414), 1.00 (404), −0.20 (446)
	ΔMAP			
BG-stm-none	0.072	(0.021, 0.123)	31-19-0	0.93 (448), 0.52 (441), −0.15 (409)
CS-stm-none	0.074	(0.034, 0.114)	38-11-1	0.39 (418), 0.35 (432), −0.30 (413)
HU-stm-none	0.083	(0.027, 0.138)	35-15-0	0.79 (441), 0.75 (414), −0.27 (421)

Table 4: Impact of Indexing All Words on GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-all-stm	−0.007	(−0.017, 0.003)	5-7-38	−0.18 (412), −0.07 (434), 0.07 (420)
CS-all-stm	0.002	(−0.004, 0.009)	2-1-47	0.10 (446), 0.09 (411), −0.07 (448)
HU-all-stm	0.002	(−0.009, 0.012)	2-1-47	0.21 (427), 0.00 (446), −0.12 (409)
	ΔMAP			
BG-all-stm	0.001	(−0.002, 0.004)	20-18-12	−0.03 (444), 0.02 (434), 0.03 (420)
CS-all-stm	−0.002	(−0.006, 0.002)	3-3-44	−0.09 (442), −0.01 (448), 0.01 (411)
HU-all-stm	−0.001	(−0.003, 0.001)	8-6-36	−0.03 (409), −0.01 (421), 0.00 (413)

tial Candidates)), for which the stemmer found apparently helpful matches such as Francouzský, francouzskou, francouzským and kandidátů. (However, we notice it did not match prezident nor prezidentem.)

3.2 Impact of Indexing All Words

Table 4 shows the impact of not discarding stopwords at index time for all 3 languages. None of the mean differences in GenS@10 were statistically significant, and few of the topics were affected.

3.3 Comparison to 4-grams

Table 5 compares the 4-gram results to stemming results for all 3 languages. None of the mean differences in GenS@10 were statistically significant, but there were large impacts on some topics in each direction. For example, for Czech topic 439 (Nehody v zaměstnání (Accidents at Work)), the 4-gram method found a relevant document with terms such as Zaměstnanci and zaměstnance that the stemmer apparently did not match.

For Hungarian, the increase in mean average precision from using 4-grams was statistically significant, presumably because Hungarian has a lot of compound words.

4 Submitted Runs

For each language, we submitted 4 experimental runs in May 2007 for official assessment. In the identifiers (e.g. "otBG07tde"), 't', 'd' and 'n' indicate that the Title, Description and Narrative field of the topic were used (respectively), and 'e' indicates that query expansion from blind feedback on the first 3 rows was used (weight of one-half on the original query, and one-sixth each on the 3 expanded rows). The 'z' code indicates that special sampling was done, as described below. From the Description and Narrative fields for most languages, instruction words such as "find", "relevant"

Table 5: 4-grams vs. Stems in GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-4gr-all	0.023	(−0.050, 0.095)	17-16-17	0.97 (414), 0.86 (415), −0.54 (411)
CS-4gr-all	−0.036	(−0.104, 0.032)	12-16-22	1.00 (439), −0.50 (409), −0.57 (429)
HU-4gr-all	0.018	(−0.055, 0.090)	14-14-22	1.00 (424), 0.63 (446), −0.94 (403)
	ΔMAP			
BG-4gr-all	0.001	(−0.044, 0.045)	26-24-0	−0.61 (445), 0.37 (417), 0.51 (414)
CS-4gr-all	0.022	(−0.015, 0.060)	29-20-1	0.40 (441), 0.37 (431), −0.34 (405)
HU-4gr-all	0.065	(0.019, 0.112)	35-15-0	0.66 (408), 0.49 (424), −0.22 (412)

Table 6: Mean Scores of Submitted Monolingual Ad Hoc Runs

Run	GS30	GS10	S10	MRR	S1	P10	GMAP	MAP
otBG07t	0.859	0.725	36/50	0.538	22/50	0.326	0.170	0.295
otBG07td	0.928	0.843	46/50	0.677	30/50	0.378	0.238	0.331
otBG07tde	0.921	0.833	44/50	0.642	27/50	0.386	0.240	0.350
(otBG07tdn)	0.937	0.858	46/50	0.680	29/50	0.398	0.257	0.351
otBG07tdnz	0.917	0.850	46/50	0.679	29/50	0.398	0.170	0.253
otCS07t	0.867	0.806	42/50	0.610	22/50	0.298	0.144	0.315
otCS07td	0.906	0.816	43/50	0.594	22/50	0.338	0.197	0.327
otCS07tde	0.894	0.805	43/50	0.623	25/50	0.356	0.191	0.348
(otCS07tdn)	0.908	0.823	43/50	0.589	20/50	0.362	0.217	0.344
otCS07tdnz	0.881	0.813	43/50	0.587	20/50	0.362	0.153	0.266
otHU07t	0.895	0.833	44/50	0.652	26/50	0.358	0.158	0.329
otHU07td	0.925	0.869	45/50	0.688	27/50	0.428	0.244	0.385
otHU07tde	0.932	0.871	45/50	0.719	30/50	0.466	0.288	0.433
(otHU07tdn)	0.939	0.887	47/50	0.712	29/50	0.444	0.281	0.411
otHU07tdnz	0.928	0.879	47/50	0.710	29/50	0.444	0.208	0.305

and "document" were automatically removed (based on looking at some older topic lists, not this year's topics; this step was skipped for Czech because we did not have an old list of Czech topics).

Details of the submitted approaches:

- "t": Just the Title field of the topic was used. Same as the "fuse" runs of Section 2.4, i.e. fusion of stemming and 4-gram runs.
- "td": Same as "t" except that the Description field was additionally used for both the stemming and 4-gram inputs.
- "tde": Same as "td" except that blind feedback (based on the first 3 rows of the "td" query) was used to expand the query. The feedback queries just used stemming, not 4-grams.
- "tdn": Same as "td" except that the Narrative field was additionally used for the stemming input (but the Narrative was still not used for the 4-gram input). (This run was not submitted.)
- "tdnz": Depth-10000 sampling run based on the "tdn" run as described below.

Table 6 lists the mean scores for the submitted runs.

Table 7: Impact of Blind Feedback on GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-tde-td	−0.010	(−0.054, 0.034)	7-13-30	0.59 (407), 0.41 (411), −0.49 (428)
CS-tde-td	−0.011	(−0.045, 0.024)	13-13-24	−0.45 (428), −0.33 (430), 0.26 (409)
HU-tde-td	0.001	(−0.024, 0.027)	10-10-30	0.38 (426), −0.16 (417), −0.33 (446)
	ΔMAP			
BG-tde-td	0.019	(−0.009, 0.047)	28-22-0	0.33 (445), 0.20 (415), −0.16 (405)
CS-tde-td	0.022	(−0.007, 0.051)	26-24-0	0.53 (433), 0.28 (431), −0.15 (441)
HU-tde-td	0.049	(0.020, 0.077)	32-17-1	0.36 (408), 0.27 (431), −0.17 (445)

4.1 Impact of Blind Feedback

Table 7 shows the impact of blind feedback on the GenS@10 and MAP measures. The results are generally consistent with our past findings that blind feedback is detrimental to GenS@10 even when it boosts MAP, though the the mean differences for GenS@10 here were not statistically significant.

4.2 Depth-10000 Sampling

The submitted tdnz run for each language was actually a depth probe run from sampling the tdn run for the language (the tdn run was not itself submitted).

The base tdn run was retrieved to depth 10000 for each topic. The first 100 rows of the submitted tdnz run contained the following rows of the base tdn run in the following order:

```
1, 2, ..., 10,
20, 30, ..., 100,
200, 300, ..., 1000,
2000, 3000, ..., 10000,
15, 25, ..., 95,
150, 250, ..., 950,
1500, 2500, ..., 9500,
125, 175, ..., 975,
1250, 1750, ..., 9750.
```

The remainder of the tdnz run was the leftover rows from the base tdn run until 1000 had been retrieved (rows 11, 12, 13, 14, 16, ..., 962).

This ordering (e.g. depth 10000 before depth 15) was chosen because of uncertainty of how deep the judging would be. As long as the top-37 were judged, we would have sampling to depth 10000. The extra sample points would just improve the accuracy. The tdnz run was given highest precedence for judging. It turned out the top-60 were judged for each topic for Czech, and the top-80 were judged for each topic for Bulgarian and Hungarian.

Tables 8, 9 and 10 show the results of the sampling for each language. The columns are as follows:

- "Depth Range": The range of depths being sampled. The 11 depth ranges cover from 1 to 10000.
- "Samples": The depths of the sample points from the depth range. The samples are always uniformly spaced. They always end at the last point of the depth range. The total number of sample points (over the 11 rows of the table) adds to 60 for Czech and 80 for Bulgarian and Hungarian.

Table 8: Marginal Precision of Bulgarian Base-TDN Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	107R, 143N, 0U	0.428	1	2.1
6-10	6, 7, ..., 10	92R, 158N, 0U	0.368	1	1.8
11-50	15, 20, ..., 50	70R, 330N, 0U	0.175	5	7.0
51-100	55, 60, ..., 100	28R, 472N, 0U	0.056	5	2.8
101-200	125, 150, ..., 200	5R, 195N, 0U	0.025	25	2.5
201-500	225, 250, ..., 500	2R, 598N, 0U	0.003	25	1.0
501-900	525, 550, ..., 900	2R, 798N, 0U	0.003	25	1.0
901-1000	950, 1000	1R, 99N, 0U	0.010	50	1.0
1001-3000	1500, 2000, ..., 3000	1R, 199N, 0U	0.005	500	10.0
3001-6000	3500, 4000, ..., 6000	0R, 300N, 0U	0.000	500	0.0
6001-10000	6500, 7000, ..., 10000	0R, 400N, 0U	0.000	500	0.0

Table 9: Marginal Precision of Czech Base-TDN Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	110R, 140N, 0U	0.440	1	2.2
6-10	6, 7, ..., 10	71R, 179N, 0U	0.284	1	1.4
11-50	15, 20, ..., 50	48R, 352N, 0U	0.120	5	4.8
51-100	55, 60, ..., 100	10R, 490N, 0U	0.020	5	1.0
101-200	150, 200	3R, 97N, 0U	0.030	50	3.0
201-500	250, 300, ..., 500	1R, 299N, 0U	0.003	50	1.0
501-900	550, 600, ..., 900	3R, 397N, 0U	0.007	50	3.0
901-1000	950, 1000	1R, 99N, 0U	0.010	50	1.0
1001-3000	1500, 2000, ..., 3000	0R, 200N, 0U	0.000	500	0.0
3001-6000	3500, 4000, ..., 6000	1R, 299N, 0U	0.003	500	10.0
6001-10000	7000, 8000, ..., 10000	0R, 200N, 0U	0.000	1000	0.0

- "# Rel": The number of each type of item retrieved from the sample points over the 50 topics. The item type codes are R (relevant), N (non-relevant) and U (unjudged, of which there are always 0). The sum of the item type counts is always 50 times the number of sample points for the depth range (because there are 50 topics for each language).

- "Precision": Estimated precision of the depth range (R/(R+N+U)).

- "Wgt": The weight of each sample point. The weight is equal to the difference in ranks between sample points, i.e. each sample point can be thought of as representing this number of rows, which is itself plus the preceding unsampled rows. The weights are higher in some cases for Czech than for Bulgarian and Hungarian because we have fewer sample points for Czech (60 instead of 80).

- "EstRel/Topic": Estimated number of relevant items retrieved per topic for this depth range. This is the Precision multiplied by the size of the depth range. Or equivalently, it is (R * Wgt) / 50.

Because each sample point is at the deep end of the range of rows it represents, the sampling should tend to underestimate precision for each depth range (assuming that precision tends to fall with depth, which appears to be the case for all 3 languages).

Table 11 shows the sums of the estimated number of relevant items per topic over all depth ranges in its first row. The official number of relevant items per topic for each language is listed

Table 10: Marginal Precision of Hungarian Base-TDN Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	133R, 117N, 0U	0.532	1	2.7
6-10	6, 7, ..., 10	89R, 161N, 0U	0.356	1	1.8
11-50	15, 20, ..., 50	55R, 345N, 0U	0.138	5	5.5
51-100	55, 60, ..., 100	25R, 475N, 0U	0.050	5	2.5
101-200	125, 150, ..., 200	3R, 197N, 0U	0.015	25	1.5
201-500	225, 250, ..., 500	12R, 588N, 0U	0.020	25	6.0
501-900	525, 550, ..., 900	2R, 798N, 0U	0.003	25	1.0
901-1000	950, 1000	1R, 99N, 0U	0.010	50	1.0
1001-3000	1500, 2000, ..., 3000	0R, 200N, 0U	0.000	500	0.0
3001-6000	3500, 4000, ..., 6000	0R, 300N, 0U	0.000	500	0.0
6001-10000	6500, 7000, ..., 10000	0R, 400N, 0U	0.000	500	0.0

Table 11: Estimated Percentage of Relevant Items that are Judged, Per Topic

	BG	CS	HU
Estimated Rel@10000	29.3	27.4	21.9
Official Rel/Topic	20.2	15.2	18.2
Percentage Judged	69%	55%	83%

in the second row. The final row of the table just divides the official number of relevant items by the estimated number in the first 10000 retrieved (e.g. for Bulgarian, 20.2/29.3=69%). This number should tend to be 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 10000 rows.

However, the sampling was very coarse at the deeper ranks, e.g. for Czech, 1 relevant item out of 300 samples in the 3001-6000 range led to an estimate of 10 relevant items per topic in this range. If the sampling had turned up 0 or 2 relevant items, a minor difference, the estimate would have been 0 or 20 relevant items per topic in this range, leading to a substantially different sum (17.4 or 37.4 instead of 27.4). We should compute confidence intervals for these estimates, but have not yet done so. Also, there is a lot of variance across topics, which we have not yet analyzed.

These preliminary estimates of judging coverage for the CLEF 2007 collections (55% for Czech, 69% for Bulgarian, 83% for Hungarian) are 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) [11]. They are similar to the estimates we produced for the NTCIR-6 collections (58% for Chinese, 78% for Japanese, 100% for Korean) [13].

These incompleteness results are similar to what [15] 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, [15] 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 CLEF test collections over the years, such as from the topic analyses in [12].)

References

[1] Samir Abdou and Jacques Savoy. Considérations sur l’évaluation de la robustesse en recherche d’information. *CORIA 2007.*

[2] Harr Chen and David R. Karger. Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. *SIGIR 2006*, pp. 429-436.

[3] Cross-Language Evaluation Forum web site. http://www.clef-campaign.org/

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

[5] NTCIR (NII-NACSIS Test Collection for IR Systems) Home Page. http://research.nii.ac.jp/∼ntcadm/index-en.html

[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] Jacques Savoy. CLEF and Multilingual information retrieval resource page. http://www.unine.ch/info/clef/

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

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

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

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

[12] Stephen Tomlinson. European Ad Hoc Retrieval Experiments with Hummingbird SearchServer$^{\text{TM}}$ at CLEF 2005. *Working Notes for the CLEF 2005 Workshop.*

[13] Stephen Tomlinson. Sampling Precision to Depth 9000: Evaluation Experiments at NTCIR-6. *Proceedings of NTCIR-6*, 2007.

[14] Ellen M. Voorhees. Overview of the TREC 2004 Robust Retrieval Track. *Proceedings of TREC 2004.*

[15] Justin Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments? *SIGIR'98*, pp. 307-314.