Comparing the Robustness of Expansion Techniques and Retrieval Measures

Stephen Tomlinson
Hummingbird
Ottawa, Ontario, Canada
stephen.tomlinson@hummingbird.com
http://www.hummingbird.com/

August 20, 2006

Abstract

Hummingbird participated in the monolingual (Bulgarian, French, Hungarian, Portuguese and English) and robust (Dutch, English, French, German, Italian and Spanish) information retrieval tasks of the Ad-Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2006. In all 22 of our experiments with blind feedback (a technique known to impair robustness across topics), the mean scores of the Average Precision, Geometric MAP and Precision@10 measures increased (and most of these increases were statistically significant), implying that these measures are not suitable as robust retrieval measures. In contrast, we found that measures based on just the first relevant item, such as a Generalized Success@10 measure, successfully discerned some robustness gains, particularly the robustness advantage of expanding Title queries by using the Description field instead of blind feedback.

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

Robust Retrieval, Blind Feedback, First Relevant Score

1 Introduction

Hummingbird SearchServer[1] is a toolkit for developing enterprise search and retrieval applications. The SearchServer kernel is also embedded in other Hummingbird products for the enterprise.

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 (CLEF [2], NTCIR [4] and TREC [7]) have provided judged test collections for objective experimentation with SearchServer in more than a dozen languages.

Table 1: Sizes of CLEF 2006 Ad-Hoc Track Test Collections

Language	Text Size (uncompressed)	Documents	Topics	Rel/Topic
Portuguese	591,987,753 bytes	210,734	50	53 (lo 2, med 39, hi 266)
French	508,863,606 bytes	177,452	49	44 (lo 1, med 20, hi 521)
Bulgarian	216,432,023 bytes	69,195	50	25 (lo 2, med 15, hi 158)
Hungarian	106,631,823 bytes	49,530	48	27 (lo 4, med 17, hi 134)
English	601,737,745 bytes	169,477	49	26 (lo 1, med 17, hi 118)

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 2006 Ad-Hoc Track test collections.

2 Methodology

2.1 Data

The CLEF 2006 Ad-Hoc Track document sets consisted of tagged (SGML-formatted) news articles in 5 different languages: Bulgarian, French, Hungarian, Portuguese and English. Table 1 gives the sizes.

The CLEF organizers created 50 natural language "topics" and translated them into many languages. Some topics were 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, median 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 [11]. Accents were not indexed except for the combining breve in Bulgarian. The apostrophe was treated as a word separator for the investigated languages (except English). The custom text reader, cTREC, was updated to maintain support for the CLEF guidelines of only indexing specifically tagged fields.

Some stop words were excluded from indexing (e.g. "the", "by" and "of" in English). For these experiments, the stop word lists for Bulgarian and Hungarian were based on Savoy's updated lists [6].

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 English topic 279 whose Title was "Swiss referendums", a corresponding SearchSQL query would be:

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM CLEF06EN
WHERE FT_TEXT CONTAINS 'Swiss'|'referendums'
ORDER BY REL DESC;
```

Most aspects of the SearchServer relevance value calculation are the same as described last year [11]. Briefly, SearchServer dampens the term frequency and adjusts for document length in a manner similar to Okapi [5] 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 '2:3' and RELEVANCE_DLEN_IMP was set to 750 for all experiments in this paper.

2.4 Experimental Runs

For each language, we executed 5 experimental runs in May 2006, though just 3 were allowed to be submitted for official assessment. In the identifiers (e.g. "humBG06tde"), '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). From the Description and Narrative fields for most languages, instruction words such as "find", "relevant" and "document" were automatically removed (based on looking at some older topic lists, not this year's topics; this step was skipped for Hungarian because we did not update our lists based on last year's topics). All runs used inflections and/or derivations from stemming.

The 5 executed runs for each language:

- "t": Just the Title field of the topic was used.
- "te": Same as "t" except that blind feedback (based on the first 3 rows of the "t" query) was used to expand the query. (This run was not submitted.)
- "td": Same as "t" except that the Description field was additionally used.
- "tde": Same as "td" except that blind feedback (based on the first 3 rows of the "td" query) was used to expand the query.
- "tdn": Same as "td" except that the Narrative field was additionally used. (This run was not submitted.)

3 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. Last year, we began investigating a new measure, Generalized Success@10 (GS10) (introduced as "First Relevant Score" (FRS) in [11]), 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 a downside, namely that it pushes down the first relevant item (on average), as has now been verified not just for our own blind feedback approach, but on 7 other major blind feedback systems [9].

3.1 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. (This is an experimental new measure introduced in this paper; compared to GS10 (defined below), it further deemphasizes small differences at the top of the list.)
- *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 $r > 30$.)

3.2 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 [13]) is the primary measure for the "robust task" this year. It 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 [9] that primary recall measures better reflect robustness than GMAP.)
- *GMAP'*: We also define a linearized log average precision measure (denoted GMAP') which linearly maps the 'log average precision' values to the [0,1] interval. For statistical significance purposes, GMAP' gives the same results as GMAP, and it has advantages such as that the individual topic differences are in the familiar -1.0 to 1.0 range and are on the same scale

Table 2: Mapping of AP and GMAP'

AP	log(max(AP,0.00001))	GMAP'
0.00000	-11.51293	0.00000
0.00001	-11.51293	0.00000
0.00003	-10.36163	0.10000
0.00010	-9.21034	0.20000
0.00032	-8.05905	0.30000
0.00100	-6.90776	0.40000
0.00316	-5.75646	0.50000
0.01000	-4.60517	0.60000
0.03162	-3.45388	0.70000
0.10000	-2.30259	0.80000
0.20000	-1.60944	0.86021
0.30000	-1.20397	0.89542
0.31623	-1.15129	0.90000
0.40000	-0.91629	0.92041
0.50000	-0.69315	0.93979
0.60000	-0.51083	0.95563
0.70000	-0.35667	0.96902
0.80000	-0.22314	0.98062
0.90000	-0.10536	0.99085
1.00000	0.00000	1.00000

as the mean. Table 2 shows examples of the mapping of the AP and GMAP' scores for a topic; for example, the table shows that for GMAP, an AP increase from 0.00001 to 0.01 is considered more important than an increase from 0.01 to 1.0 (these are differences of 0.6 and 0.4 respectively in GMAP'). (This example illustrates one of our concerns with GMAP, which is that small differences likely to be unimportant to a user can be dramatically amplified.)

3.3 Statistical Significance Tables

For tables comparing 2 diagnostic runs (such as Table 4), 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-td-t" specifies the difference of subtracting the scores of the Bulgarian 't' run from the Bulgarian 'td' run (of Table 3).
- "ΔGS30" is the difference of the mean GS30 scores of the two runs being compared (and "ΔGS10" is the difference of the mean GS10 scores, etc.).
- "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

Table 3: Mean Scores of Monolingual Ad Hoc Runs

Run	GS30	GS10	S10	MRR	S1	P10	GMAP	MAP
(humBG06tdn)	0.919	0.846	46/50	0.651	26/50	0.322	0.200	0.305
humBG06td	0.903	0.829	44/50	0.648	26/50	0.308	0.172	0.285
humBG06tde	0.896	0.826	44/50	0.616	23/50	0.334	0.182	0.305
(humBG06te)	0.858	0.742	38/50	0.537	21/50	0.314	0.148	0.291
humBG06t	0.832	0.724	38/50	0.513	19/50	0.282	0.108	0.261
(humFR06tdn)	0.980	0.946	48/49	0.833	37/49	0.465	0.332	0.427
humFR06td	0.968	0.921	47/49	0.781	33/49	0.445	0.294	0.387
humFR06tde	0.962	0.910	47/49	0.759	32/49	0.486	0.317	0.416
(humFR06te)	0.940	0.861	45/49	0.674	28/49	0.433	0.266	0.390
humFR06t	0.936	0.857	44/49	0.702	30/49	0.390	0.233	0.352
(humHU06tdn)	0.949	0.882	43/48	0.735	31/48	0.360	0.197	0.293
humHU06td	0.954	0.881	45/48	0.675	26/48	0.377	0.203	0.298
humHU06tde	0.966	0.906	46/48	0.693	26/48	0.423	0.240	0.336
(humHU06te)	0.923	0.820	42/48	0.590	21/48	0.392	0.188	0.309
humHU06t	0.911	0.817	43/48	0.582	20/48	0.354	0.128	0.267
(humPT06tdn)	0.957	0.912	49/50	0.737	31/50	0.584	0.299	0.420
humPT06td	0.961	0.910	49/50	0.739	30/50	0.574	0.314	0.426
humPT06tde	0.957	0.913	48/50	0.764	33/50	0.602	0.337	0.451
(humPT06te)	0.927	0.874	46/50	0.700	29/50	0.568	0.248	0.420
humPT06t	0.928	0.872	46/50	0.695	28/50	0.542	0.220	0.391
(humEN06tdn)	0.958	0.896	45/49	0.791	36/49	0.453	0.336	0.445
humEN06td	0.957	0.885	46/49	0.683	27/49	0.404	0.306	0.409
humEN06tde	0.947	0.868	45/49	0.668	26/49	0.451	0.325	0.449
(humEN06te)	0.912	0.815	41/49	0.633	25/49	0.396	0.258	0.404
humEN06t	0.910	0.814	41/49	0.643	26/49	0.378	0.229	0.371

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).

4 Results of Query Expansion Experiments

4.1 Expansion of Title Queries

Table 4 shows that expanding the Title queries by adding the Description field increased the mean score for all investigated measures (GS30, GS10, MRR, P10, GMAP and MAP), including at least one statistically significant increase for each measure. Adding the Description is a "robust" technique that can sometimes improve a poor result from just using the Title field.

Table 5 shows that expanding the Title queries via blind feedback of the first 3 rows did not produce any statistically significant increases for the primary recall measures (GS30, GS10, MRR), even though it produced statistically significant increases for the secondary recall measures (P10, GMAP, MAP). Blind feedback is not a robust technique in that it is unlikely to improve poor results. (In a larger experiment, we would expect the primary recall measures to show statistically significant decreases, like we saw for Bulgarian last year [11].)

Table 6 compares the results of the two title-expansion approaches. For each primary recall measure (GS30, GS10, MRR), there is at least one positive statistically significant difference,

Table 4: Impact of Adding the Description to Title-only Queries

Expt	ΔGS30	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-td-t	0.071	(0.010, 0.133)	22-7-21	0.78 (308), 0.76 (357), −0.64 (322)
FR-td-t	0.032	(−0.001, 0.064)	12-9-28	0.58 (309), 0.43 (325), −0.06 (347)
HU-td-t	0.044	(−0.004, 0.091)	15-9-24	0.85 (309), 0.52 (358), −0.25 (369)
PT-td-t	0.032	(0.000, 0.064)	17-9-24	0.49 (327), 0.49 (346), −0.08 (344)
EN-td-t	0.046	(0.012, 0.081)	18-9-22	0.53 (338), 0.43 (343), −0.09 (346)
	ΔGS10			
BG-td-t	0.105	(0.024, 0.186)	22-7-21	0.99 (308), 0.99 (357), −0.62 (322)
FR-td-t	0.064	(0.003, 0.125)	12-9-28	0.94 (309), 0.84 (325), −0.16 (347)
HU-td-t	0.065	(−0.005, 0.135)	15-9-24	0.91 (358), 0.63 (321), −0.42 (369)
PT-td-t	0.038	(−0.003, 0.079)	17-9-24	0.59 (346), 0.48 (327), −0.21 (344)
EN-td-t	0.071	(0.015, 0.127)	18-9-22	0.91 (338), 0.54 (349), −0.23 (346)
	ΔMRR			
BG-td-t	0.135	(0.038, 0.232)	22-7-21	0.98 (308), 0.98 (357), −0.67 (303)
FR-td-t	0.078	(−0.023, 0.180)	12-9-28	0.97 (309), 0.96 (325), −0.50 (318)
HU-td-t	0.092	(−0.004, 0.189)	15-9-24	0.97 (358), 0.89 (365), −0.75 (320)
PT-td-t	0.044	(−0.052, 0.140)	17-9-24	0.75 (318), 0.75 (350), −0.75 (303)
EN-td-t	0.040	(−0.056, 0.136)	18-9-22	0.97 (338), 0.91 (349), −0.67 (307)
	ΔP10			
BG-td-t	0.026	(−0.026, 0.078)	18-13-19	0.50 (361), 0.50 (364), −0.50 (314)
FR-td-t	0.055	(0.006, 0.104)	18-9-22	0.40 (328), 0.40 (303), −0.40 (316)
HU-td-t	0.023	(−0.027, 0.073)	19-8-21	−0.60 (315), 0.40 (354), 0.40 (372)
PT-td-t	0.032	(−0.022, 0.086)	19-13-18	0.80 (337), 0.40 (323), −0.40 (304)
EN-td-t	0.027	(−0.018, 0.071)	15-13-21	0.50 (338), 0.40 (349), −0.20 (305)
	ΔGMAP'			
BG-td-t	0.041	(0.009, 0.072)	34-15-1	0.44 (324), 0.43 (308), −0.13 (315)
FR-td-t	0.020	(0.003, 0.037)	30-19-0	0.21 (317), 0.20 (309), −0.11 (315)
HU-td-t	0.040	(0.006, 0.074)	36-9-3	0.72 (309), 0.32 (358), −0.16 (315)
PT-td-t	0.031	(0.001, 0.061)	30-19-1	0.59 (320), 0.34 (323), −0.12 (315)
EN-td-t	0.025	(0.010, 0.040)	25-21-3	0.22 (338), 0.13 (313), −0.04 (324)
	ΔMAP			
BG-td-t	0.024	(−0.015, 0.063)	34-15-1	−0.35 (373), −0.31 (315), 0.32 (353)
FR-td-t	0.034	(0.001, 0.068)	30-19-0	−0.28 (315), 0.25 (341), 0.27 (311)
HU-td-t	0.031	(0.009, 0.054)	36-9-3	0.22 (318), −0.19 (353), −0.20 (315)
PT-td-t	0.035	(0.000, 0.070)	30-19-1	0.46 (341), 0.35 (337), −0.32 (315)
EN-td-t	0.038	(−0.002, 0.079)	25-21-3	0.66 (341), 0.44 (338), −0.25 (324)

Table 5: Impact of Blind Feedback Expansion on Title-only Queries

Expt	ΔGS30	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-te-t	0.026	(−0.012, 0.065)	14-11-25	0.83 (324), 0.34 (374), −0.17 (357)
FR-te-t	0.004	(−0.010, 0.017)	7-8-34	0.19 (325), 0.14 (320), −0.09 (336)
HU-te-t	0.012	(−0.022, 0.046)	8-9-31	0.77 (309), −0.11 (374), −0.12 (357)
PT-te-t	−0.001	(−0.012, 0.009)	6-8-36	0.17 (323), −0.05 (343), −0.16 (327)
EN-te-t	0.002	(−0.015, 0.018)	11-7-31	−0.21 (322), 0.14 (343), 0.19 (309)
	ΔGS10			
BG-te-t	0.018	(−0.011, 0.047)	14-11-25	0.54 (324), 0.28 (374), −0.16 (352)
FR-te-t	0.005	(−0.020, 0.029)	7-8-34	0.24 (325), 0.22 (317), −0.21 (347)
HU-te-t	0.003	(−0.022, 0.028)	8-9-31	0.43 (309), −0.16 (375), −0.23 (357)
PT-te-t	0.002	(−0.015, 0.018)	6-8-36	0.25 (323), −0.11 (325), −0.14 (343)
EN-te-t	0.001	(−0.022, 0.023)	11-7-31	0.26 (309), −0.18 (322), −0.25 (318)
	ΔMRR			
BG-te-t	0.024	(−0.018, 0.067)	14-11-25	0.50 (360), 0.50 (363), −0.50 (365)
FR-te-t	−0.029	(−0.082, 0.025)	7-8-34	−0.67 (340), −0.67 (312), 0.50 (328)
HU-te-t	0.007	(−0.041, 0.056)	8-9-31	−0.50 (314), −0.50 (364), 0.50 (372)
PT-te-t	0.006	(−0.051, 0.063)	6-8-36	−0.67 (343), −0.50 (326), 0.50 (312)
EN-te-t	−0.010	(−0.061, 0.040)	11-7-31	−0.67 (345), 0.50 (329), 0.50 (325)
	ΔP10			
BG-te-t	0.032	(0.006, 0.058)	13-4-33	0.40 (364), 0.30 (319), −0.10 (304)
FR-te-t	0.043	(0.006, 0.080)	19-7-23	0.40 (345), 0.40 (328), −0.20 (346)
HU-te-t	0.038	(0.008, 0.067)	13-3-32	0.40 (319), 0.30 (354), −0.20 (305)
PT-te-t	0.026	(−0.006, 0.058)	17-5-28	−0.40 (316), −0.20 (331), 0.30 (342)
EN-te-t	0.018	(−0.012, 0.049)	15-4-30	0.30 (328), −0.30 (305), −0.30 (344)
	ΔGMAP'			
BG-te-t	0.028	(−0.004, 0.059)	35-14-1	0.78 (324), 0.08 (310), −0.04 (368)
FR-te-t	0.012	(0.006, 0.017)	42-7-0	0.06 (323), 0.04 (307), −0.05 (312)
HU-te-t	0.034	(0.000, 0.067)	40-5-3	0.78 (309), 0.10 (368), −0.02 (357)
PT-te-t	0.010	(0.003, 0.017)	38-10-2	0.10 (323), 0.08 (337), −0.04 (327)
EN-te-t	0.010	(0.003, 0.018)	37-9-3	0.12 (309), 0.05 (341), −0.04 (316)
	ΔMAP			
BG-te-t	0.029	(0.018, 0.041)	35-14-1	0.15 (364), 0.14 (319), −0.02 (373)
FR-te-t	0.038	(0.021, 0.055)	42-7-0	0.25 (341), 0.12 (328), −0.14 (312)
HU-te-t	0.042	(0.025, 0.059)	40-5-3	0.28 (319), 0.18 (364), −0.01 (321)
PT-te-t	0.029	(0.014, 0.044)	38-10-2	0.15 (341), 0.13 (306), −0.14 (332)
EN-te-t	0.033	(0.018, 0.049)	37-9-3	0.20 (336), 0.19 (341), −0.10 (344)

Table 6: Comparison of Expansion Techniques for Title-only Queries

Expt	ΔGS30	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-td-te	0.045	(−0.021, 0.112)	18-10-22	0.93 (357), 0.80 (308), −0.65 (324)
FR-td-te	0.028	(−0.002, 0.059)	15-8-26	0.64 (309), 0.25 (325), −0.10 (320)
HU-td-te	0.032	(−0.002, 0.065)	15-8-25	0.42 (358), 0.35 (352), −0.27 (369)
PT-td-te	0.034	(−0.002, 0.069)	18-10-22	0.65 (327), 0.51 (346), −0.07 (303)
EN-td-te	0.045	(0.010, 0.080)	16-9-24	0.49 (338), 0.38 (322), −0.11 (346)
	ΔGS10			
BG-td-te	0.087	(0.004, 0.170)	18-10-22	1.00 (357), 0.99 (308), −0.54 (324)
FR-td-te	0.060	(0.005, 0.114)	15-8-26	0.96 (309), 0.60 (325), −0.17 (320)
HU-td-te	0.062	(−0.005, 0.128)	15-8-25	0.83 (358), 0.69 (321), −0.47 (369)
PT-td-te	0.037	(−0.003, 0.076)	18-10-22	0.59 (346), 0.50 (327), −0.21 (303)
EN-td-te	0.070	(0.012, 0.128)	16-9-24	0.88 (338), 0.51 (313), −0.30 (346)
	ΔMRR			
BG-td-te	0.111	(0.007, 0.214)	18-10-22	0.99 (357), 0.99 (308), −0.67 (303)
FR-td-te	0.107	(0.001, 0.212)	15-8-26	0.98 (309), 0.92 (325), −0.50 (337)
HU-td-te	0.085	(−0.022, 0.192)	15-8-25	0.96 (358), 0.90 (365), −0.75 (372)
PT-td-te	0.038	(−0.061, 0.137)	18-10-22	0.75 (350), 0.75 (318), −0.75 (303)
EN-td-te	0.051	(−0.037, 0.139)	16-9-24	0.97 (338), 0.90 (349), −0.67 (307)
	ΔP10			
BG-td-te	−0.006	(−0.058, 0.046)	15-17-18	−0.50 (314), −0.40 (373), 0.50 (361)
FR-td-te	0.012	(−0.038, 0.063)	15-15-19	−0.50 (345), 0.40 (303), 0.40 (340)
HU-td-te	−0.015	(−0.068, 0.039)	16-14-18	−0.60 (315), −0.60 (311), 0.40 (372)
PT-td-te	0.006	(−0.049, 0.061)	13-17-20	0.60 (316), 0.60 (337), −0.30 (336)
EN-td-te	0.008	(−0.037, 0.053)	14-19-16	0.50 (338), 0.30 (344), −0.20 (328)
	ΔGMAP'			
BG-td-te	0.013	(−0.016, 0.042)	25-25-0	0.37 (308), 0.26 (357), −0.34 (324)
FR-td-te	0.009	(−0.006, 0.024)	22-25-2	0.18 (317), 0.18 (309), −0.11 (315)
HU-td-te	0.007	(−0.013, 0.026)	23-22-3	0.30 (358), −0.13 (368), −0.18 (315)
PT-td-te	0.021	(−0.008, 0.049)	23-25-2	0.59 (320), 0.24 (323), −0.13 (315)
EN-td-te	0.015	(−0.001, 0.030)	26-20-3	0.24 (338), 0.13 (313), −0.07 (336)
	ΔMAP			
BG-td-te	−0.006	(−0.041, 0.030)	25-25-0	−0.33 (373), −0.31 (315), 0.30 (353)
FR-td-te	−0.004	(−0.037, 0.030)	22-25-2	−0.28 (315), −0.27 (345), 0.23 (317)
HU-td-te	−0.011	(−0.040, 0.018)	23-22-3	−0.29 (311), −0.27 (315), 0.16 (324)
PT-td-te	0.006	(−0.029, 0.041)	23-25-2	−0.35 (315), −0.34 (313), 0.31 (341)
EN-td-te	0.005	(−0.034, 0.045)	26-20-3	0.47 (341), 0.45 (338), −0.30 (324)

reflecting the robustness of using the Description instead of blind feedback. However, there are no statistically significant differences in the secondary recall measures (P10, GMAP, MAP); these measures do not discern the higher robustness of the "td" run compared to the "te" run.

4.2 Expansion of "Title+Desc" Queries

Table 7 shows that expanding the Description queries by adding the Narrative field tended to be beneficial for both primary and secondary recall measures, though not as consistently as was adding the Description to the Title queries. (Sometimes the Narrative field specifies what is not relevant.)

Table 8 produced a lot of statistically significant increases for the secondary recall measures (P10, GMAP, MAP). We also see one statistically significant increase for a primary recall measure (for Hungarian), which we suspect is a Type I error, because it does not fit the pattern we have seen over several other experiments [11, 8, 10, 9] (including last year's Hungarian experiment, for which mean GS10 was down slightly with blind feedback [11]).

Table 9 compares the results of the two expansion approaches for "Title+Desc" queries. The Narrative was modestly beneficial for the primary recall measures compared to blind feedback, reflecting a robustness advantage, even though blind feedback boosted the secondary recall measures a little more.

Table 7: Impact of Adding the Narrative to "Title+Desc" Queries

Expt	ΔGS30	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-tdn-td	0.016	(−0.012, 0.045)	13-12-25	0.45 (320), 0.43 (358), −0.17 (315)
FR-tdn-td	0.012	(−0.008, 0.033)	11-6-32	0.49 (336), 0.05 (320), −0.05 (312)
HU-tdn-td	−0.005	(−0.037, 0.026)	16-9-23	−0.64 (362), −0.20 (352), 0.14 (357)
PT-tdn-td	−0.004	(−0.020, 0.012)	11-11-28	−0.26 (320), −0.13 (345), 0.13 (323)
EN-tdn-td	0.001	(−0.018, 0.021)	16-6-27	−0.33 (318), −0.15 (310), 0.11 (331)
	ΔGS10			
BG-tdn-td	0.017	(−0.024, 0.058)	13-12-25	0.66 (320), 0.30 (374), −0.43 (315)
FR-tdn-td	0.025	(−0.014, 0.064)	11-6-32	0.88 (336), 0.13 (323), −0.13 (312)
HU-tdn-td	0.000	(−0.055, 0.056)	16-9-23	−0.96 (362), −0.43 (352), 0.30 (375)
PT-tdn-td	0.001	(−0.030, 0.033)	11-11-28	0.37 (323), −0.26 (318), −0.37 (345)
EN-tdn-td	0.011	(−0.034, 0.056)	16-6-27	−0.64 (318), −0.42 (310), 0.32 (331)
	ΔMRR			
BG-tdn-td	0.004	(−0.070, 0.077)	13-12-25	0.67 (303), −0.67 (306), −0.67 (313)
FR-tdn-td	0.053	(−0.022, 0.127)	11-6-32	0.97 (336), 0.50 (338), −0.50 (335)
HU-tdn-td	0.061	(−0.054, 0.175)	16-9-23	−0.98 (362), −0.83 (365), 0.75 (320)
PT-tdn-td	−0.002	(−0.097, 0.093)	11-11-28	0.86 (323), −0.80 (318), −0.86 (345)
EN-tdn-td	0.107	(0.014, 0.201)	16-6-27	−0.88 (310), 0.83 (313), 0.83 (331)
	ΔP10			
BG-tdn-td	0.014	(−0.023, 0.051)	16-11-23	−0.30 (313), 0.30 (314), 0.30 (363)
FR-tdn-td	0.020	(−0.023, 0.064)	14-13-22	0.70 (331), 0.40 (316), −0.30 (312)
HU-tdn-td	−0.017	(−0.072, 0.039)	14-18-16	−0.80 (362), −0.50 (313), 0.50 (311)
PT-tdn-td	0.010	(−0.025, 0.045)	19-12-19	−0.40 (313), −0.30 (316), 0.20 (322)
EN-tdn-td	0.049	(0.012, 0.085)	19-6-24	0.40 (316), 0.40 (331), −0.20 (350)
	ΔGMAP'			
BG-tdn-td	0.013	(−0.004, 0.029)	31-17-2	0.24 (320), 0.21 (358), −0.17 (315)
FR-tdn-td	0.011	(−0.003, 0.024)	30-18-1	0.29 (336), 0.07 (331), −0.07 (312)
HU-tdn-td	−0.002	(−0.011, 0.007)	25-23-0	−0.09 (362), −0.08 (305), 0.04 (311)
PT-tdn-td	−0.004	(−0.017, 0.009)	19-29-2	−0.27 (320), 0.04 (323), 0.08 (327)
EN-tdn-td	0.008	(−0.001, 0.018)	31-15-3	0.10 (313), 0.09 (334), −0.08 (318)
	ΔMAP			
BG-tdn-td	0.020	(0.000, 0.041)	31-17-2	0.18 (320), 0.16 (357), −0.16 (306)
FR-tdn-td	0.040	(−0.004, 0.084)	30-18-1	0.97 (336), 0.26 (331), −0.16 (312)
HU-tdn-td	−0.005	(−0.034, 0.024)	25-23-0	−0.27 (362), −0.23 (318), 0.25 (301)
PT-tdn-td	−0.006	(−0.028, 0.015)	19-29-2	−0.33 (334), 0.13 (342), 0.15 (331)
EN-tdn-td	0.037	(0.007, 0.066)	31-15-3	0.32 (340), 0.32 (334), −0.15 (344)

Table 8: Impact of Blind Feedback Expansion on "Title+Desc" Queries

Expt	ΔGS30	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-tde-td	−0.007	(−0.025, 0.011)	9-14-27	−0.35 (320), −0.11 (322), 0.11 (310)
FR-tde-td	−0.006	(−0.013, 0.001)	7-11-31	−0.08 (320), −0.07 (336), 0.02 (322)
HU-tde-td	0.012	(0.000, 0.023)	12-4-32	0.22 (374), 0.10 (369), −0.03 (357)
PT-tde-td	−0.003	(−0.014, 0.007)	10-9-31	−0.19 (320), −0.09 (327), 0.08 (344)
EN-tde-td	−0.010	(−0.029, 0.008)	7-11-31	−0.31 (343), −0.18 (316), 0.20 (309)
	ΔGS10			
BG-tde-td	−0.002	(−0.022, 0.018)	9-14-27	0.23 (354), 0.19 (310), −0.14 (306)
FR-tde-td	−0.011	(−0.026, 0.005)	7-11-31	−0.14 (303), −0.14 (331), 0.07 (322)
HU-tde-td	0.025	(0.002, 0.047)	12-4-32	0.38 (374), 0.21 (363), −0.07 (364)
PT-tde-td	0.002	(−0.014, 0.019)	10-9-31	0.21 (344), −0.11 (303), −0.16 (327)
EN-tde-td	−0.017	(−0.049, 0.015)	7-11-31	−0.40 (316), −0.35 (343), 0.39 (309)
	ΔMRR			
BG-tde-td	−0.032	(−0.096, 0.032)	9-14-27	−0.67 (306), −0.50 (351), 0.50 (363)
FR-tde-td	−0.022	(−0.103, 0.059)	7-11-31	−0.67 (313), −0.67 (331), 0.50 (322)
HU-tde-td	0.019	(−0.035, 0.073)	12-4-32	0.50 (322), 0.50 (314), −0.50 (364)
PT-tde-td	0.025	(−0.037, 0.087)	10-9-31	0.50 (324), 0.50 (312), −0.50 (326)
EN-tde-td	−0.016	(−0.068, 0.037)	7-11-31	−0.67 (327), −0.50 (311), 0.50 (324)
	ΔP10			
BG-tde-td	0.026	(−0.004, 0.056)	11-6-33	0.50 (315), 0.30 (312), −0.10 (368)
FR-tde-td	0.041	(0.009, 0.073)	20-7-22	0.50 (349), 0.20 (335), −0.20 (322)
HU-tde-td	0.046	(0.008, 0.084)	18-7-23	0.40 (319), 0.30 (364), −0.30 (372)
PT-tde-td	0.028	(−0.003, 0.059)	21-9-20	0.20 (315), 0.20 (329), −0.20 (305)
EN-tde-td	0.047	(0.012, 0.081)	20-8-21	0.30 (345), 0.30 (349), −0.20 (316)
	ΔGMAP'			
BG-tde-td	0.005	(−0.003, 0.013)	30-19-1	0.07 (310), 0.06 (315), −0.06 (306)
FR-tde-td	0.007	(0.002, 0.012)	31-16-2	0.04 (307), 0.04 (349), −0.04 (320)
HU-tde-td	0.015	(0.008, 0.022)	35-11-2	0.07 (309), 0.06 (361), −0.04 (371)
PT-tde-td	0.006	(0.001, 0.010)	33-15-2	0.05 (313), 0.04 (337), −0.03 (305)
EN-tde-td	0.005	(−0.002, 0.013)	34-12-3	−0.09 (343), 0.05 (336), 0.05 (309)
	ΔMAP			
BG-tde-td	0.020	(0.002, 0.038)	30-19-1	−0.22 (306), 0.15 (364), 0.17 (301)
FR-tde-td	0.029	(0.011, 0.047)	31-16-2	0.24 (349), 0.16 (340), −0.14 (327)
HU-tde-td	0.037	(0.019, 0.055)	35-11-2	0.19 (319), 0.17 (362), −0.07 (304)
PT-tde-td	0.024	(0.009, 0.040)	33-15-2	0.19 (337), 0.15 (313), −0.16 (305)
EN-tde-td	0.040	(0.018, 0.062)	34-12-3	0.26 (338), 0.20 (349), −0.17 (327)

Table 9: Comparison of Expansion Techniques for "Title+Desc" Queries

Expt	ΔGS30	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-tdn-tde	0.023	(−0.016, 0.063)	17-12-21	0.81 (320), 0.43 (358), −0.20 (324)
FR-tdn-tde	0.018	(−0.006, 0.042)	12-4-33	0.55 (336), 0.13 (320), −0.07 (312)
HU-tdn-tde	−0.017	(−0.049, 0.015)	12-12-24	−0.64 (362), −0.20 (352), 0.17 (357)
PT-tdn-tde	−0.001	(−0.014, 0.012)	10-13-27	0.17 (327), 0.11 (323), −0.13 (345)
EN-tdn-tde	0.012	(−0.017, 0.040)	17-5-27	0.37 (343), 0.25 (316), −0.33 (318)
	ΔGS10			
BG-tdn-tde	0.019	(−0.024, 0.063)	17-12-21	0.73 (320), 0.34 (311), −0.43 (315)
FR-tdn-tde	0.035	(−0.006, 0.077)	12-4-33	0.93 (336), 0.17 (320), −0.21 (312)
HU-tdn-tde	−0.024	(−0.081, 0.032)	12-12-24	−0.96 (362), −0.43 (352), 0.34 (357)
PT-tdn-tde	−0.001	(−0.034, 0.032)	10-13-27	−0.37 (345), 0.32 (323), 0.34 (327)
EN-tdn-tde	0.028	(−0.029, 0.085)	17-5-27	−0.64 (318), 0.46 (343), 0.60 (316)
	ΔMRR			
BG-tdn-tde	0.035	(−0.050, 0.121)	17-12-21	−0.75 (309), −0.67 (313), 0.67 (303)
FR-tdn-tde	0.075	(−0.003, 0.152)	12-4-33	0.97 (336), 0.67 (303), −0.75 (312)
HU-tdn-tde	0.042	(−0.066, 0.150)	12-12-24	−0.98 (362), −0.83 (365), 0.75 (320)
PT-tdn-tde	−0.027	(−0.124, 0.071)	10-13-27	−0.86 (345), −0.67 (315), 0.83 (323)
EN-tdn-tde	0.123	(0.031, 0.215)	17-5-27	0.92 (316), 0.86 (313), −0.50 (305)
	ΔP10			
BG-tdn-tde	−0.012	(−0.056, 0.032)	14-14-22	−0.60 (315), −0.30 (312), 0.30 (314)
FR-tdn-tde	−0.020	(−0.066, 0.025)	8-13-28	−0.60 (349), −0.30 (340), 0.60 (331)
HU-tdn-tde	−0.063	(−0.118,−0.007)	12-22-14	−0.90 (362), −0.40 (319), 0.20 (354)
PT-tdn-tde	−0.018	(−0.051, 0.015)	12-14-24	−0.50 (313), −0.20 (329), 0.20 (322)
EN-tdn-tde	0.002	(−0.048, 0.053)	11-15-23	0.60 (316), 0.50 (331), −0.40 (345)
	ΔGMAP'			
BG-tdn-tde	0.008	(−0.011, 0.028)	25-24-1	0.27 (320), 0.23 (358), −0.24 (315)
FR-tdn-tde	0.004	(−0.011, 0.019)	20-28-1	0.31 (336), 0.07 (331), −0.06 (312)
HU-tdn-tde	−0.017	(−0.028,−0.006)	18-30-0	−0.11 (362), −0.10 (309), 0.04 (320)
PT-tdn-tde	−0.010	(−0.023, 0.003)	15-34-1	−0.26 (320), −0.09 (313), 0.09 (327)
EN-tdn-tde	0.003	(−0.009, 0.015)	19-26-4	0.12 (343), 0.11 (313), −0.09 (345)
	ΔMAP			
BG-tdn-tde	0.001	(−0.023, 0.025)	25-24-1	0.20 (357), 0.18 (320), −0.19 (315)
FR-tdn-tde	0.011	(−0.036, 0.057)	20-28-1	0.97 (336), 0.26 (331), −0.25 (349)
HU-tdn-tde	−0.043	(−0.074,−0.012)	18-30-0	−0.44 (362), −0.21 (318), 0.17 (301)
PT-tdn-tde	−0.031	(−0.054,−0.007)	15-34-1	−0.33 (334), −0.20 (313), 0.15 (331)
EN-tdn-tde	−0.003	(−0.036, 0.029)	19-26-4	0.30 (331), 0.28 (334), −0.30 (345)

Table 10: Mean Scores of Robust "Training" and "Test" Runs

Run	GS30	GS10	S10	MRR	S1	P10	GMAP	MAP
humDE06Rtd0	0.963	0.923	59/60	0.747	37/60	0.530	0.322	0.422
humDE06Rtde0	0.966	0.932	59/60	0.786	41/60	0.553	0.367	0.470
humEN06Rtd0	0.887	0.795	46/55	0.660	32/55	0.342	0.204	0.385
humEN06Rtde0	0.887	0.789	45/55	0.666	33/55	0.378	0.225	0.424
humES06Rtd0	0.947	0.896	55/59	0.761	40/59	0.464	0.263	0.367
humES06Rtde0	0.924	0.871	53/59	0.729	37/59	0.485	0.285	0.407
humFR06Rtd0	0.908	0.858	54/60	0.702	35/60	0.337	0.241	0.403
humFR06Rtde0	0.912	0.856	54/60	0.682	33/60	0.357	0.269	0.432
humIT06Rtd0	0.915	0.825	53/57	0.571	23/57	0.360	0.209	0.357
humIT06Rtde0	0.910	0.817	53/57	0.556	22/57	0.389	0.232	0.392
humNL06Rtd0	0.959	0.902	56/60	0.776	42/60	0.492	0.342	0.438
humNL06Rtde0	0.953	0.901	57/60	0.757	39/60	0.503	0.373	0.475
humDE06Rtd	0.964	0.904	91/95	0.706	54/95	0.483	0.330	0.447
humDE06Rtde	0.953	0.891	91/95	0.711	56/95	0.539	0.382	0.508
humEN06Rtd	0.952	0.899	82/88	0.737	54/88	0.382	0.373	0.486
humEN06Rtde	0.952	0.900	82/88	0.760	58/88	0.435	0.419	0.541
humES06Rtd	0.945	0.879	90/97	0.702	56/97	0.473	0.302	0.433
humES06Rtde	0.917	0.850	86/97	0.706	59/97	0.500	0.322	0.471
humFR06Rtd	0.944	0.888	86/91	0.709	53/91	0.412	0.337	0.470
humFR06Rtde	0.951	0.895	85/91	0.732	57/91	0.430	0.386	0.499
humIT06Rtd	0.940	0.885	84/90	0.676	46/90	0.407	0.281	0.409
humIT06Rtde	0.945	0.892	85/90	0.692	48/90	0.447	0.324	0.466
humNL06Rtd	0.954	0.907	92/96	0.785	68/96	0.478	0.346	0.484
humNL06Rtde	0.948	0.899	89/96	0.779	67/96	0.531	0.393	0.532

Table 11: Impact of Blind Feedback on Robust "Training" Topics

Expt	ΔGS30	95% Conf	vs.	3 Extreme Diffs (Topic)
DE-e0	0.003	(−0.003, 0.008)	12-6-42	0.08 (151), 0.05 (51), −0.05 (58)
EN-e0	−0.000	(−0.030, 0.029)	7-9-39	0.56 (56), −0.23 (52), −0.34 (124)
ES-e0	−0.024	(−0.046,−0.001)	2-12-45	−0.52 (57), −0.29 (185), 0.02 (104)
FR-e0	0.004	(−0.008, 0.015)	10-7-43	0.20 (109), 0.10 (101), −0.16 (125)
IT-e0	−0.005	(−0.019, 0.010)	7-12-38	−0.29 (108), 0.11 (104), 0.18 (101)
NL-e0	−0.005	(−0.026, 0.015)	7-11-42	−0.40 (183), −0.13 (106), 0.38 (126)
	ΔGS10			
DE-e0	0.009	(−0.004, 0.022)	12-6-42	0.18 (151), 0.14 (51), −0.14 (58)
EN-e0	−0.005	(−0.037, 0.026)	7-9-39	0.49 (56), −0.28 (120), −0.48 (124)
ES-e0	−0.025	(−0.047,−0.004)	2-12-45	−0.37 (101), −0.34 (185), 0.07 (104)
FR-e0	−0.002	(−0.023, 0.019)	10-7-43	−0.37 (125), 0.20 (101), 0.21 (109)
IT-e0	−0.008	(−0.031, 0.015)	7-12-38	−0.44 (108), −0.16 (185), 0.34 (101)
NL-e0	−0.000	(−0.027, 0.027)	7-11-42	0.61 (126), −0.21 (54), −0.26 (183)
	ΔMRR			
DE-e0	0.039	(−0.014, 0.091)	12-6-42	0.67 (51), 0.50 (125), −0.67 (58)
EN-e0	0.006	(−0.014, 0.026)	7-9-39	0.50 (74), −0.09 (120), −0.09 (124)
ES-e0	−0.033	(−0.074, 0.009)	2-12-45	−0.67 (105), −0.50 (59), 0.50 (104)
FR-e0	−0.020	(−0.062, 0.022)	10-7-43	−0.67 (186), −0.67 (108), 0.50 (76)
IT-e0	−0.015	(−0.048, 0.018)	7-12-38	−0.50 (76), −0.50 (58), 0.50 (55)
NL-e0	−0.019	(−0.059, 0.021)	7-11-42	−0.50 (159), −0.50 (104), 0.50 (184)
	ΔP10			
DE-e0	0.023	(−0.009, 0.055)	19-11-30	0.50 (57), 0.30 (78), −0.20 (58)
EN-e0	0.036	(0.006, 0.067)	16-5-34	0.50 (159), 0.40 (107), −0.20 (124)
ES-e0	0.020	(−0.009, 0.050)	16-14-29	0.40 (52), 0.30 (151), −0.20 (73)
FR-e0	0.020	(−0.009, 0.049)	19-8-33	0.30 (122), −0.30 (73), −0.30 (127)
IT-e0	0.030	(0.008, 0.052)	21-8-28	0.20 (70), 0.20 (187), −0.10 (129)
NL-e0	0.012	(−0.018, 0.042)	14-13-33	0.50 (154), 0.30 (156), −0.20 (104)
	ΔGMAP'			
DE-e0	0.011	(0.005, 0.017)	42-16-2	0.10 (51), 0.05 (128), −0.07 (54)
EN-e0	0.008	(0.001, 0.016)	35-13-7	0.08 (159), 0.08 (56), −0.07 (52)
ES-e0	0.007	(0.000, 0.013)	33-25-1	0.08 (156), 0.07 (52), −0.05 (57)
FR-e0	0.009	(0.004, 0.015)	38-15-7	0.08 (129), 0.05 (105), −0.04 (73)
IT-e0	0.009	(0.004, 0.014)	41-11-5	0.09 (101), 0.04 (183), −0.04 (185)
NL-e0	0.008	(0.001, 0.014)	42-16-2	0.09 (126), 0.06 (51), −0.05 (54)
	ΔMAP			
DE-e0	0.048	(0.030, 0.066)	42-16-2	0.31 (57), 0.23 (159), −0.05 (58)
EN-e0	0.039	(0.015, 0.063)	35-13-7	0.39 (159), 0.30 (107), −0.10 (55)
ES-e0	0.040	(0.018, 0.061)	33-25-1	0.28 (52), 0.27 (156), −0.11 (79)
FR-e0	0.029	(0.013, 0.046)	38-15-7	0.19 (122), 0.15 (153), −0.17 (123)
IT-e0	0.035	(0.022, 0.048)	41-11-5	0.15 (126), 0.14 (74), −0.05 (185)
NL-e0	0.037	(0.016, 0.059)	42-16-2	0.33 (154), 0.21 (156), −0.16 (188)

Table 12: Impact of Blind Feedback on Robust "Test" Topics

Expt	ΔGS30	95% Conf	vs.	3 Extreme Diffs (Topic)
DE-e1	−0.011	(−0.023, 0.002)	10-18-67	−0.49 (161), −0.20 (137), 0.09 (95)
EN-e1	−0.000	(−0.008, 0.008)	13-8-67	−0.23 (178), −0.14 (139), 0.12 (113)
ES-e1	−0.028	(−0.059, 0.002)	16-18-63	−0.84 (114), −0.78 (139), 0.17 (111)
FR-e1	0.006	(−0.003, 0.016)	17-11-63	0.34 (200), 0.13 (111), −0.10 (148)
IT-e1	0.005	(−0.004, 0.014)	18-14-58	0.26 (47), 0.21 (69), −0.06 (139)
NL-e1	−0.005	(−0.025, 0.014)	11-15-70	0.58 (195), −0.22 (169), −0.58 (113)
	ΔGS10			
DE-e1	−0.013	(−0.028, 0.003)	10-18-67	−0.36 (137), −0.36 (161), 0.17 (95)
EN-e1	0.001	(−0.013, 0.014)	13-8-67	−0.44 (178), 0.10 (168), 0.17 (169)
ES-e1	−0.029	(−0.058, 0.000)	16-18-63	−0.73 (114), −0.73 (96), 0.21 (47)
FR-e1	0.007	(−0.005, 0.019)	17-11-63	0.21 (46), 0.14 (91), −0.20 (148)
IT-e1	0.008	(−0.007, 0.023)	18-14-58	0.42 (69), 0.14 (118), −0.14 (143)
NL-e1	−0.008	(−0.030, 0.014)	11-15-70	−0.66 (113), −0.40 (164), 0.42 (195)
	ΔMRR			
DE-e1	0.006	(−0.033, 0.044)	10-18-67	−0.67 (48), 0.67 (139), 0.67 (111)
EN-e1	0.022	(−0.011, 0.056)	13-8-67	0.50 (171), 0.50 (196), −0.50 (115)
ES-e1	0.005	(−0.039, 0.048)	16-18-63	−0.67 (43), −0.50 (147), 0.67 (164)
FR-e1	0.024	(−0.019, 0.067)	17-11-63	0.67 (91), 0.50 (133), −0.50 (179)
IT-e1	0.016	(−0.035, 0.066)	18-14-58	0.67 (118), 0.67 (117), −0.67 (143)
NL-e1	−0.006	(−0.046, 0.033)	11-15-70	−0.67 (172), 0.67 (117), 0.67 (116)
	ΔP10			
DE-e1	0.056	(0.029, 0.082)	41-9-45	0.50 (200), 0.40 (117), −0.40 (190)
EN-e1	0.053	(0.026, 0.081)	33-9-46	0.50 (140), 0.50 (86), −0.20 (193)
ES-e1	0.027	(0.007, 0.046)	31-12-54	0.30 (143), 0.20 (113), −0.20 (96)
FR-e1	0.018	(−0.002, 0.037)	26-16-49	0.30 (192), 0.20 (63), −0.20 (163)
IT-e1	0.040	(0.016, 0.064)	34-10-46	−0.40 (143), 0.40 (192), 0.40 (141)
NL-e1	0.053	(0.028, 0.079)	39-14-43	0.30 (133), 0.30 (132), −0.30 (147)
	ΔGMAP'			
DE-e1	0.013	(0.006, 0.019)	76-15-4	−0.13 (161), 0.09 (111), 0.12 (95)
EN-e1	0.010	(0.005, 0.015)	56-19-13	−0.06 (178), 0.06 (196), 0.06 (64)
ES-e1	0.006	(0.000, 0.011)	66-28-3	−0.11 (166), 0.06 (99), 0.09 (67)
FR-e1	0.012	(0.006, 0.017)	65-17-9	0.09 (46), 0.08 (200), −0.06 (84)
IT-e1	0.012	(0.007, 0.018)	69-17-4	0.10 (118), 0.07 (192), −0.09 (166)
NL-e1	0.011	(0.004, 0.018)	71-17-8	0.16 (195), 0.12 (88), −0.11 (113)
	ΔMAP			
DE-e1	0.062	(0.044, 0.079)	76-15-4	0.34 (200), 0.33 (99), −0.09 (137)
EN-e1	0.055	(0.031, 0.079)	56-19-13	0.50 (196), 0.36 (192), −0.18 (43)
ES-e1	0.038	(0.022, 0.054)	66-28-3	0.34 (192), 0.27 (200), −0.14 (112)
FR-e1	0.030	(0.006, 0.054)	65-17-9	−0.50 (84), −0.50 (141), 0.23 (198)
IT-e1	0.057	(0.034, 0.080)	69-17-4	−0.50 (165), 0.30 (118), 0.44 (192)
NL-e1	0.048	(0.029, 0.067)	71-17-8	0.32 (114), 0.30 (179), −0.19 (147)

5 Robust Task Results

The "Robust Task" re-used the old test collections for Dutch, English, French, German, Italian and Spanish from CLEF 2001-2003. Of the 160 old topics, 60 were allowed to be used for new "training", leaving the other 100 for "testing". Participants were encouraged to train on the GMAP measure, though we believe primary recall measures better reflect robustness. We actually did not do any new training for this task.

Note that even though the document sets were not always the same for each language in 2001, 2002 and 2003, a fixed document set was used for each language in this task. Hence there may be more unjudged relevant items than usual. Unfortunately, we did not have time to look at metrics on just judged items for this paper.

Table 10 lists the mean scores of our submitted Robust Task runs. For each language, we submitted a "td" run (no blind feedback) and a "tde" run (incorporating blind feedback based on the first 3 rows of "td"). Even though blind feedback is known to tend to make results less robust, the GMAP score was higher with blind feedback in all cases (as were P10 and MAP).

Tables 11 and 12 isolate the impact of blind feedback on each measure. The impact on the primary recall measures tended to be detrimental, including a statistically significant decrease on the Spanish training topics. The increases on the secondary recall measures were mostly statistically significant. While this generally fits the pattern we have seen in other experiments (e.g. [9]), the negative impact on the primary recall measures seems to be less strong than we have seen elsewhere. Perhaps the old CLEF topics tend to be "easier" than, say, the old TREC topics used at RIA [9], providing relatively fewer cases for which blind feedback would be detrimental.

6 Conclusions

For all 22 blind feedback experiments reported in this paper, the mean scores for MAP, GMAP and P10 were up with blind feedback, and most of these increases were statistically significant. As blind feedback is known to be bad for robustness (because of its tendency to "not help (and frequently hurt) the worst performing topics" [12]), we conclude that none of these 3 measures should be used as robustness measures.

Measures based on just the first relevant item (i.e. primary recall measures such as GS30 and GS10) reflect robustness. In this paper, we found in particular that these measures discerned the robustness advantage of expanding Title queries by using the Description field instead of blind feedback, while the secondary recall measures (MAP, GMAP, P10) did not.

These results are consistent with what we have seen elsewhere [11, 8, 10, 9]. For example, in [9], 7 other groups' blind feedback systems were studied, and it was found that blind feedback was detrimental to the first relevant item (on average), even though it boosted the secondary recall measures.

A paper at the recent SIGIR conference [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.

To encourage more research in robust retrieval, probably the simplest thing the organizers of ad hoc tracks could do would be to use a measure based on just the first relevant item (e.g. GS10 or GS30) as the primary measure for the ad hoc task. Participants would then find it detrimental to use the non-robust blind feedback technique, but potentially would be rewarded for finding ways of producing more diverse results.

References

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

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

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

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

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

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

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

[8] Stephen Tomlinson. CJK Experiments with Hummingbird SearchServer™ at NTCIR-5. *Proceedings of NTCIR-5*, 2005.

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

[10] Stephen Tomlinson. Enterprise, QA, Robust and Terabyte Experiments with Hummingbird SearchServer™ at TREC 2005. *Proceedings of TREC 2005.*

[11] Stephen Tomlinson. European Ad Hoc Retrieval Experiments with Hummingbird SearchServer™ at CLEF 2005. *Working Notes for the CLEF 2005 Workshop.*

[12] Ellen M. Voorhees. Overview of the TREC 2003 Robust Retrieval Track. *Proceedings of TREC 2003.*

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