German, French, English and Persian Retrieval Experiments at CLEF 2009

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

August 23, 2009

Abstract

We describe evaluation experiments conducted by submitting retrieval runs for the monolingual German, French, English and Persian (Farsi) information retrieval tasks of the Ad Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2009. In the ad hoc retrieval tasks, the system was given 50 natural language queries, and the goal was to find all of the relevant records or documents (with high precision) in a particular document set. We conducted diagnostic experiments with different techniques for matching word variations, 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 decompounding and stemming, but generally disagreed on the blind feedback technique. 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 German, French and English and 80 for Persian. The results suggest that, on average, the percentage of relevant items assessed was less than 62% for German, 27% for French, 35% for English and 22% for Persian.

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

German Retrieval, French Retrieval, English Retrieval, Persian Retrieval, Robust Retrieval, Sampling

1 Introduction

Open Text eDOCS SearchServer^{TM} is a toolkit for developing enterprise search and retrieval applications. The eDOCS SearchServer kernel is also embedded in various components of the Open Text eDOCS Suite[1].

[1] Open Text eDOCS SearchServer and Open Text eDOCS Suite 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 2009 Ad Hoc Track Test Collections

Code	Language	Text Size (uncompressed)	Documents	Topics	Rel/Topic
DE	German	1,306,492,248 bytes	869,353	50	31 (lo 3, hi 86)
EN	English	1,208,383,351 bytes	1,000,100	50	51 (lo 8, hi 235)
FA	Persian	628,471,252 bytes	166,774	50	89 (lo 8, hi 266)
FR	French	1,362,122,091 bytes	1,000,100	50	37 (lo 2, hi 120)

The eDOCS SearchServer kernel 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 [9]) have provided judged test collections for objective experimentation with the SearchServer kernel in more than a dozen languages.

This paper describes experimental work with SearchServer (experimental post-6.0 versions) for the task of finding relevant documents for natural language queries in various European languages using the CLEF 2009 Ad Hoc Track test collections.

2 Methodology

2.1 Data

The CLEF 2009 Ad Hoc Track document sets were the same as used in 2008. They consisted of XML-tagged records or documents in 4 different languages: German, French, English and Persian (also known as Farsi). For German, French and English, the records were library catalog cards (bibliographic records describing publications archived by The European Library (TEL)). For Persian, the documents were newspaper articles (Hamshahri corpus of 1996-2002). Table 1 gives the collection sizes.

The CLEF organizers created 50 natural language "topics" (numbered 701-750 for German, French and English and 601-650 for Persian) and translated them into many languages. Sometimes topics are discarded for some languages because of a lack of relevant documents (though that did not happen this year). 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, please see the track overview paper.

2.2 Indexing

Our indexing approach was the same as last year [18]. Accents were not indexed. The apostrophe was treated as a word separator (except in English). The custom text reader, cTREC, enforced the CLEF guidelines by only marking specifically tagged fields for indexing.

For some experiments, some stop words were excluded from indexing (e.g. words like "the", "by" and "of" in English). For our Persian experiments, our stop word list was based on Savoy's list [8].

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 German topic 701 whose Title was "Tiere in der Arktis" (Arctic Animals) a corresponding SearchSQL query would be:

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM CLEF09DE
WHERE FT_TEXT CONTAINS 'Tiere'|'in'|'der'|'Arktis'
ORDER BY REL DESC;
```

Most aspects of the SearchServer relevance value calculation are the same as described last year [18]. Briefly, SearchServer dampens the term frequency and adjusts for document length in a manner similar to Okapi [7] 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/decompound' was previously specified). The SearchServer RELEVANCE_METHOD setting was set to '2:3' and RELEVANCE_DLEN_IMP was set to 500 for all experiments in this paper.

2.4 Diagnostic Runs

For the diagnostic runs listed in Table 2, the run names consist of a language code ("DE" for German, "EN" for English, "FA" for Persian, and "FR" for French) 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).
- "lexstem" (German, French and English only): Same as "none" except that linguistic variations from stemming were matched. The lexicon-based inflectional stemmer in SearchServer was used. For German, this stemmer includes decompounding.
- "algstem": Same as "lexstem" except that an algorithmic stemmer was used. For Persian, our stemmer was ported from Savoy's [8]. For German, French and English, Porter's algorithmic "Snowball" stemmers [6] were used (for English, the Porter2 version was used).
- "lexall" (German, French and English only): Same as "lexstem" except that a separate index was used which did not stop any words from being indexed.
- "algall" (Persian only): Same as "algstem" except that a separate index was used which did not stop any words from being indexed.
- "4gram": 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 or word stopping 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.

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 relevant documents, 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. In 2005, we began investigating a new measure, Generalized Success@10 (GenS@10 or GS10) (introduced as "First Relevant Score" (FRS) in [13]), 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 [11] 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* (GenS@30 or 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* (GenS@10 or 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 relevant documents found) to 1.0 (all relevant documents 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 [19]) 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 [11] that primary recall measures better reflect robustness than GMAP.)

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. (We abbreviate "lexstem" to "lex", "algstem" to "alg", "4gram" to "4gr" and "lexall" and "algall" to "all".) The difference is the first run minus the second run. For example, "DE-lex-none" specifies the difference of subtracting the scores of the German 'none' run from the German 'lexstem' 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 4 languages. For instance, it shows that the mean increase in GenS@10 was statistically significant for German, and the mean increases in MAP were statistically significant for all 4 languages.

Table 3 also shows that there were large impacts from stemming on particular topics for German, French and English in both the GenS@10 and MAP measures (we look at some examples in the later sections).

Like last year, for Persian, even on individual topics there was relatively little impact from stemming. We notice in Table 2 that the Success@10 rate was relatively high for Persian (50 out of 50) even without stemming, and that relevant documents were plentiful (89 per topic on average as per Table 1), but we have not done sufficient analysis to understand why the stemming impact was so minor across Persian topics.

Table 2: Mean Scores of Diagnostic Monolingual Ad Hoc Runs

Run	GS30	GS10	S10	MRR	S1	P10	GMAP	MAP
DE-lexall	0.932	0.865	45/50	0.690	29/50	0.350	0.158	0.246
DE-lexstem	0.929	0.858	46/50	0.665	27/50	0.370	0.168	0.255
DE-4gram	0.918	0.825	42/50	0.622	25/50	0.324	0.118	0.218
DE-algstem	0.750	0.693	38/50	0.534	22/50	0.288	0.023	0.173
DE-none	0.679	0.609	32/50	0.482	20/50	0.232	0.011	0.133
EN-lexstem	0.943	0.867	46/50	0.693	29/50	0.418	0.178	0.286
EN-lexall	0.937	0.861	46/50	0.698	30/50	0.420	0.174	0.281
EN-algstem	0.938	0.855	44/50	0.649	25/50	0.434	0.180	0.282
EN-none	0.913	0.836	46/50	0.648	26/50	0.394	0.124	0.246
EN-4gram	0.888	0.821	42/50	0.648	26/50	0.396	0.086	0.229
FA-algall	0.987	0.961	50/50	0.850	39/50	0.604	0.316	0.384
FA-4gram	0.987	0.961	50/50	0.829	36/50	0.602	0.308	0.369
FA-none	0.984	0.951	50/50	0.805	35/50	0.584	0.298	0.365
FA-algstem	0.983	0.950	50/50	0.822	37/50	0.584	0.311	0.382
FR-lexstem	0.812	0.733	39/50	0.555	21/50	0.356	0.080	0.239
FR-algstem	0.802	0.731	39/50	0.547	20/50	0.336	0.087	0.235
FR-lexall	0.801	0.715	37/50	0.508	17/50	0.318	0.065	0.211
FR-4gram	0.802	0.713	38/50	0.499	17/50	0.282	0.051	0.174
FR-none	0.782	0.696	38/50	0.532	21/50	0.322	0.062	0.204

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

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
DE-lex-none	0.249	(0.129, 0.369)	23-10-17	1.00 (708), 1.00 (743), −0.37 (733)
FR-lex-none	0.037	(−0.036, 0.111)	9-16-25	1.00 (704), 1.00 (731), −0.46 (733)
EN-lex-none	0.031	(−0.014, 0.075)	9-8-33	0.79 (729), 0.50 (710), −0.14 (727)
FA-alg-none	−0.001	(−0.025, 0.023)	8-9-33	0.27 (628), −0.21 (645), −0.25 (646)
	ΔMAP			
DE-lex-none	0.121	(0.074, 0.168)	41-9-0	0.63 (708), 0.51 (720), −0.22 (733)
FR-lex-none	0.035	(0.010, 0.061)	25-22-3	0.35 (744), 0.27 (713), −0.09 (722)
EN-lex-none	0.040	(0.010, 0.071)	29-18-3	0.42 (744), 0.31 (709), −0.19 (742)
FA-alg-none	0.017	(0.003, 0.032)	26-23-1	0.17 (644), 0.15 (650), −0.06 (610)

Table 4: Lexical vs. Algorithmic Stemming in GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
DE-lex-alg	0.165	(0.062, 0.268)	19-11-20	1.00 (743), 1.00 (738), −0.29 (736)
EN-lex-alg	0.012	(−0.003, 0.027)	10-1-39	−0.21 (735), 0.14 (739), 0.14 (746)
FR-lex-alg	0.002	(−0.054, 0.058)	6-6-38	1.00 (718), −0.36 (750), −0.77 (717)
	ΔMAP			
DE-lex-alg	0.081	(0.035, 0.128)	35-14-1	0.63 (708), 0.51 (720), −0.41 (744)
EN-lex-alg	0.004	(−0.010, 0.017)	27-11-12	−0.23 (742), 0.11 (705), 0.17 (702)
FR-lex-alg	0.004	(−0.008, 0.016)	25-17-8	0.23 (744), 0.07 (706), −0.07 (722)

3.2 Lexical vs. Algorithmic Stemming

Table 4 isolates the differences between the lexical and algorithmic stemmers for the 3 languages for which both types of stemmers were available. For each language, each stemmer outscored the other on at least some individual topics. The higher mean scores of lexical stemming for German were statistically significant in both the GenS@10 and MAP measures.

German is a language with frequent compound words, and the lexical stemmer included decompounding, unlike the algorithmic stemmer. For example, in German topic 738 (Naturkundemuseen (Natural History Museums)) the run using lexical stemming and decompounding matched passages in relevant records such as 'Naturhistorisches Museum' and 'Naturhistorischen Museums' that were missed by the algorithmic stemmer. The algorithmic stemmer also missed the variant of 'Naturkundemuseum' which was common in relevant records.

Unfortunately, we haven't had time to walk through more of the stemming differences, but in the past we found a lot of them were from the lexical stemmers just matching inflections while the algorithmic stemmers often additionally match derivations [15].

3.3 Impact of Stop Words

Table 5 shows the impact of using stop words for the 4 languages. Occasionally we see a surprisingly large impact from using stop words on individual topics. For example, for French topic 704 (Bienfaits sociaux du sport (Social Benefits of Sport)) the relevant record found at rank 1 by the base run (oai:bnf.fr:catalogue/ark:/12148/cb37099470n/description) fell to rank 11 when the noise word "du" was not stopped, partly because it did not contain the word "du" (or any linguistic variant of "du"). While "du" had less weight than the other query terms from a lower inverse document frequency (idf), the term occurred only in 217,669 of the 1,000,100 records, and so it apparently still had enough weight to influence the results.

While the mean impact of stopping tended to be small, the mean increases in MAP were statistically significant in French and German.

3.4 Comparison to 4-grams

Table 6 compares the 4-gram results to stemming results for all 4 languages. (The 4-gram index did not stop any words from being indexed, so the comparison is to the stemming runs which likewise did not use stop words.) While most of the mean differences were not statistically significant, there were a lot of large differences on individual topics.

For example, in German topic 739 (Ozonabbau (Ozone Depletion)), the lexical stemming run substantially outscored the 4-gram run in the GenS@10 measure. The stemmer produced stems of 'ozon' and 'abbau', with 'ozon' getting a little higher weight from inverse document frequency. The relevant records typically just used the 'ozon' stem, and relevant records were retrieved as high as rank 3. While the 4-gram approach also produced the 4-character 'ozon' as a search term, its matches were filled with records with surnames containing 'nabbau' which had three times the

Table 5: Impact of Stop Words on GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
FR-lex-all	0.018	(−0.016, 0.052)	10-7-33	0.54 (704), 0.39 (732), −0.35 (717)
EN-lex-all	0.006	(−0.012, 0.024)	3-2-45	0.39 (732), 0.07 (747), −0.14 (715)
DE-lex-all	−0.007	(−0.022, 0.008)	3-4-43	−0.21 (711), −0.19 (748), 0.13 (732)
FA-alg-all	−0.011	(−0.025, 0.003)	0-4-46	−0.32 (646), −0.07 (609), 0.00 (626)
	ΔMAP			
FR-lex-all	0.029	(0.015, 0.042)	31-8-11	0.16 (721), 0.15 (703), −0.03 (742)
EN-lex-all	0.005	(−0.007, 0.018)	11-6-33	0.23 (741), −0.07 (721), −0.10 (715)
DE-lex-all	0.009	(0.001, 0.017)	18-6-26	0.11 (741), 0.09 (740), −0.03 (748)
FA-alg-all	−0.001	(−0.010, 0.007)	11-8-31	0.12 (643), −0.10 (609), −0.10 (646)

Table 6: Stems vs. 4-grams in GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
EN-all-4gr	0.040	(−0.011, 0.091)	15-8-27	1.00 (711), 0.58 (723), −0.19 (737)
DE-all-4gr	0.040	(−0.033, 0.112)	15-12-23	0.86 (739), 0.78 (747), −0.50 (726)
FR-all-4gr	0.002	(−0.087, 0.092)	16-13-21	−1.00 (702), −1.00 (710), 0.88 (714)
FA-all-4gr	−0.001	(−0.025, 0.023)	10-7-33	−0.32 (635), −0.21 (645), 0.27 (628)
	ΔMAP			
EN-all-4gr	0.052	(0.017, 0.087)	39-10-1	0.58 (706), 0.31 (711), −0.23 (742)
DE-all-4gr	0.028	(−0.012, 0.068)	28-22-0	0.45 (745), −0.35 (717), −0.40 (707)
FR-all-4gr	0.037	(−0.003, 0.076)	32-17-1	−0.51 (702), 0.34 (713), 0.36 (744)
FA-all-4gr	0.015	(0.000, 0.029)	29-21-0	0.17 (644), 0.15 (650), −0.06 (646)

weight from three 4-grams (‘nabb’, ‘abba’, ‘bbau’) and it did not retrieve a relevant record until rank 509.

4 Submitted Runs

For each language, we submitted 4 experimental runs in June 2009 for official assessment. In the identifiers (e.g. “otFA09tdz”), ‘t’ and ‘d’ indicate that the Title and Description 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 field for German, French and English, 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 Persian as we didn’t have time to update our lists this year).

Details of the submitted approaches:

- “t”: Just the Title field of the topic was used. Same as the “lexstem” run of Section 2.4 for German, French and English, and same as the “algstem” run of Section 2.4 for Persian.
- “td”: Same as “t” except that the Description field of the topic 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.
- “tdz”: Depth-10000 sampling run based on the “td” run as described below.

Table 7 lists the mean scores for the submitted runs.

Table 7: Mean Scores of Submitted Monolingual Ad Hoc Runs

Run	GS30	GS10	S10	MRR	S1	P10	GMAP	MAP
otDE09t	0.929	0.858	46/50	0.665	27/50	0.370	0.168	0.255
otDE09td	0.925	0.851	46/50	0.641	25/50	0.364	0.171	0.257
otDE09tde	0.903	0.812	43/50	0.599	23/50	0.408	0.181	0.287
otDE09tdz	0.935	0.860	46/50	0.643	25/50	0.364	0.125	0.188
otEN09t	0.943	0.867	46/50	0.693	29/50	0.418	0.178	0.286
otEN09td	0.970	0.921	47/50	0.767	32/50	0.480	0.228	0.316
otEN09tde	0.964	0.914	48/50	0.745	30/50	0.510	0.246	0.346
otEN09tdz	0.966	0.919	47/50	0.767	32/50	0.480	0.159	0.216
otFA09t	0.983	0.950	50/50	0.822	37/50	0.584	0.311	0.382
otFA09td	0.984	0.951	50/50	0.771	31/50	0.578	0.314	0.374
otFA09tde	0.978	0.936	50/50	0.722	26/50	0.612	0.329	0.395
otFA09tdz	0.984	0.951	50/50	0.771	31/50	0.578	0.235	0.271
otFR09t	0.812	0.733	39/50	0.555	21/50	0.356	0.080	0.239
otFR09td	0.866	0.752	40/50	0.571	23/50	0.356	0.135	0.240
otFR09tde	0.826	0.729	40/50	0.530	20/50	0.326	0.113	0.241
otFR09tdz	0.840	0.741	40/50	0.569	23/50	0.356	0.096	0.169

Table 8: Impact of the Description Field on GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
EN-td-t	0.054	(−0.008, 0.116)	16-10-24	0.91 (749), 0.85 (704), −0.39 (737)
FR-td-t	0.019	(−0.052, 0.090)	20-10-20	−0.89 (747), 0.69 (717), 0.73 (748)
FA-td-t	0.001	(−0.028, 0.031)	7-11-32	0.39 (646), 0.25 (635), −0.21 (612)
DE-td-t	−0.008	(−0.073, 0.058)	10-14-26	0.93 (713), −0.45 (701), −0.75 (743)
	ΔMAP			
EN-td-t	0.030	(−0.014, 0.073)	29-20-1	0.43 (731), 0.39 (702), −0.39 (705)
FR-td-t	0.001	(−0.025, 0.027)	29-20-1	−0.31 (743), −0.17 (709), 0.22 (748)
FA-td-t	−0.008	(−0.031, 0.015)	21-29-0	0.34 (618), −0.13 (608), −0.15 (630)
DE-td-t	0.002	(−0.044, 0.048)	27-23-0	−0.60 (710), 0.36 (734), 0.50 (742)

4.1 Impact of Including the Description Field

Table 8 shows the impact of including the Description field on the GenS@10 and MAP measures. We see large impacts on individual topics in both directions, but none of the mean differences were statistically significant.

4.2 Impact of Blind Feedback

Table 9 shows the impact of blind feedback on the GenS@10 and MAP measures. GenS@10 declined with blind feedback for all four languages, and while none of the mean differences in GenS@10 were statistically significant, German and Persian were close to the borderline. MAP increased with blind feedback for all four languages, including statistically significant increases for German and English, and Persian was again close to the borderline. These results are generally consistent with our past findings that blind feedback is detrimental to GenS@10 even when it boosts MAP [11].

Table 9: Impact of Blind Feedback on GenS@10 and Average Precision

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
EN-tde-td	−0.007	(−0.032, 0.018)	5-7-38	−0.40 (736), −0.22 (748), 0.27 (726)
FA-tde-td	−0.015	(−0.033, 0.002)	6-12-32	−0.25 (622), −0.16 (612), 0.07 (637)
FR-tde-td	−0.023	(−0.073, 0.027)	12-19-19	0.70 (747), −0.32 (708), −0.68 (750)
DE-tde-td	−0.039	(−0.080, 0.003)	8-10-32	−0.73 (710), −0.43 (719), 0.13 (727)
	ΔMAP			
EN-tde-td	0.030	(0.013, 0.047)	32-17-1	0.20 (733), 0.17 (734), −0.11 (731)
FA-tde-td	0.021	(−0.002, 0.044)	28-22-0	0.36 (630), 0.27 (621), −0.08 (650)
FR-tde-td	0.001	(−0.016, 0.017)	23-25-2	0.14 (705), 0.14 (716), −0.13 (742)
DE-tde-td	0.030	(0.008, 0.052)	35-15-0	0.42 (716), 0.18 (749), −0.12 (723)

4.3 Depth-10000 Sampling

The submitted tdz run for each language was actually a depth probe run from sampling the td run for the language.

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

```
1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
20, 30, 40, 50, 60, 70, 80, 90, 100,
200, 300, 400, 500, 600, 700, 800, 900, 1000,
2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000,
15, 25, ..., 95,
150, 250, ..., 950,
1500, 2500, ..., 9500,
125, 175, ..., 975,
1250, 1750, ..., 9750.
```

The remainder of the sample run was padded with the top-ranked remaining rows from the base run until 1000 rows had been retrieved (i.e. rows 11, 12, 13, 14, 16, ..., 962 of the base run).

This ordering (e.g. the placement of the sample from depth 10000 before the sample from depth 15) was chosen because of uncertainty of how deep the judging would be. As long as the top-37 were judged for each topic, we would have sampling to depth 10000 (because in the above list, you can count that after 37 samples that depth 10000 is reached). The extra sample points, if judged, would just improve the accuracy (because they are just additional sample points from the top 10000, not deeper sample points).

Our sample run (i.e. the tdz run) for each language was submitted to the CLEF organizers for assessing in June 2009. We assigned it highest judging precedence of all our submitted runs.

When we received the relevance judgments and analyzed them in August 2009, we checked the judging depth of our sample runs. We found that the top-60 rows were judged for each topic for each German, French and English, and the top-80 rows were judged for each topic for Persian.

Tables 10, 11, 12 and 13 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 German, French and English and adds to 80 for Persian.

Table 10: Marginal Precision of German Base-TD Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	106R, 144N, 0U	0.424	1	2.1
6-10	6, 7, ..., 10	76R, 174N, 0U	0.304	1	1.5
11-50	15, 20, ..., 50	72R, 328N, 0U	0.180	5	7.2
51-100	55, 60, ..., 100	44R, 456N, 0U	0.088	5	4.4
101-200	150, 200	5R, 95N, 0U	0.050	50	5.0
201-500	250, 300, ..., 500	5R, 295N, 0U	0.017	50	5.0
501-900	550, 600, ..., 900	5R, 395N, 0U	0.013	50	5.0
901-1000	950, 1000	0R, 100N, 0U	0.000	50	0.0
1001-3000	1500, 2000, ..., 3000	1R, 199N, 0U	0.005	500	10.0
3001-6000	3500, 4000, ..., 6000	1R, 295N, 4X	0.003	500	10.0
6001-10000	7000, 8000, ..., 10000	0R, 189N, 11X	0.000	1000	0.0

Table 11: Marginal Precision of French Base-TD Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	90R, 160N, 0U	0.360	1	1.8
6-10	6, 7, ..., 10	88R, 162N, 0U	0.352	1	1.8
11-50	15, 20, ..., 50	89R, 311N, 0U	0.223	5	8.9
51-100	55, 60, ..., 100	60R, 440N, 0U	0.120	5	6.0
101-200	150, 200	8R, 92N, 0U	0.080	50	8.0
201-500	250, 300, ..., 500	15R, 285N, 0U	0.050	50	15.0
501-900	550, 600, ..., 900	5R, 395N, 0U	0.013	50	5.0
901-1000	950, 1000	2R, 98N, 0U	0.020	50	2.0
1001-3000	1500, 2000, ..., 3000	3R, 196N, 1X	0.015	500	30.0
3001-6000	3500, 4000, ..., 6000	4R, 288N, 8X	0.013	500	40.0
6001-10000	7000, 8000, ..., 10000	1R, 189N, 10X	0.005	1000	20.0

- "# 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). An X is used when a sample point was not submitted because fewer than 10000 rows were retrieved for the topic (which just happened for a few topics). 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+X)).

- "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.

- "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 4 languages).

Table 14 shows the sums of the estimated number of relevant items per topic over all depth ranges in its first row (i.e. it is the sum of the EstRel/Topic entries in the last column of the

Table 12: Marginal Precision of English Base-TD Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	128R, 122N, 0U	0.512	1	2.6
6-10	6, 7, ..., 10	112R, 138N, 0U	0.448	1	2.2
11-50	15, 20, ..., 50	107R, 293N, 0U	0.268	5	10.7
51-100	55, 60, ..., 100	44R, 456N, 0U	0.088	5	4.4
101-200	150, 200	7R, 93N, 0U	0.070	50	7.0
201-500	250, 300, ..., 500	12R, 288N, 0U	0.040	50	12.0
501-900	550, 600, ..., 900	15R, 385N, 0U	0.037	50	15.0
901-1000	950, 1000	2R, 98N, 0U	0.020	50	2.0
1001-3000	1500, 2000, ..., 3000	4R, 196N, 0U	0.020	500	40.0
3001-6000	3500, 4000, ..., 6000	1R, 297N, 2X	0.003	500	10.0
6001-10000	7000, 8000, ..., 10000	2R, 194N, 4X	0.010	1000	40.0

Table 13: Marginal Precision of Persian Base-TD Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	158R, 92N, 0U	0.632	1	3.2
6-10	6, 7, ..., 10	131R, 119N, 0U	0.524	1	2.6
11-50	15, 20, ..., 50	159R, 241N, 0U	0.398	5	15.9
51-100	55, 60, ..., 100	151R, 349N, 0U	0.302	5	15.1
101-200	125, 150, ..., 200	38R, 162N, 0U	0.190	25	19.0
201-500	225, 250, ..., 500	93R, 507N, 0U	0.155	25	46.5
501-900	525, 550, ..., 900	81R, 719N, 0U	0.101	25	40.5
901-1000	950, 1000	7R, 93N, 0U	0.070	50	7.0
1001-3000	1500, 2000, ..., 3000	7R, 193N, 0U	0.035	500	70.0
3001-6000	3500, 4000, ..., 6000	10R, 290N, 0U	0.033	500	100.0
6001-10000	6500, 7000, ..., 10000	9R, 388N, 3X	0.022	500	90.0

Table 14: Estimated Percentage of Relevant Items that are Judged

	DE	FR	EN	FA
Estimated Rel@10000	50.2	138.5	145.9	409.8
Official Rel/Topic	31.2	37.1	50.5	89.3
Percentage Judged	62%	27%	35%	22%

corresponding table from Tables 10-13). The official number of relevant items per topic for each language is listed 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 German, 31.2/50.2=62%). 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 French, 1 relevant item out of 200 samples in the 6001-10000 range led to an estimate of 20 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 40 relevant items per topic in this range, leading to a substantially different sum (118.5 or 158.5 instead of 138.5). 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 2009 collections are similar to last year's estimates [18] for two of the four languages (62% for German this year, 55% last year; 22% for Persian this year, 25% last year). For the other two languages this year's estimates are substantially lower than last year's (27% for French this year, 52% last year; 35% for English this year, 53% last year). We've used similar methodology (though sometimes using different sampling depths) for other past collections, such as the CLEF 2007 Ad Hoc collections (55% for Czech, 69% for Bulgarian, 83% for Hungarian) [17], the NTCIR-7 ACLIA IR4QA collections (65% for Simplified Chinese, 32% for Traditional Chinese, 41% for Japanese) [14], the NTCIR-6 CLIR collections (58% for Chinese, 78% for Japanese, 100% for Korean) [16], and the TREC 2006 Legal and Terabyte collections (18% for TREC Legal and 36% for TREC Terabyte) [12].

On older TREC collections of approximately 500,000 documents which used depth-100 pooling, [20] reported that "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, [20] also found for such test collections that "overall they do indeed lead to reliable results."

5 Conclusions

We described evaluation experiments conducted by submitting retrieval runs for the monolingual German, French, English and Persian (Farsi) information retrieval tasks of the Ad Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2009. We conducted diagnostic experiments with different techniques for matching word variations, 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 decompounding and stemming, but generally disagreed on the blind feedback technique. 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 German, French and English and 80 for Persian. The results suggest that, on average, the percentage of relevant items assessed was less than 62% for German, 27% for French, 35% for English and 22% for Persian. We analyzed a few individual topics for which the different retrieval techniques produced large differences in the scores and found that the judgments were sufficient to gain insight into the reasons for the differences.

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] M. F. Porter. Snowball: A language for stemming algorithms. October 2001. http://snowball.tartarus.org/texts/introduction.html

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

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

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

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

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

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

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

[14] Experiments in Finding Chinese and Japanese Answer Documents at NTCIR-7. *Proceedings of NTCIR-7*, 2008.

[15] Stephen Tomlinson. Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer^{TM} at CLEF 2003. *Working Notes for the CLEF 2003 Workshop.*

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

[17] Stephen Tomlinson. Sampling Precision to Depth 10000: Evaluation Experiments at CLEF 2007. *Working Notes for the CLEF 2007 Workshop.*

[18] Stephen Tomlinson. German, French, English and Persian Retrieval Experiments at CLEF 2008. *Working Notes for the CLEF 2008 Workshop.*

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

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