Experiments in 8 European Languages with Hummingbird SearchServer™ at CLEF 2002

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

August 30, 2002

Abstract

Hummingbird submitted ranked result sets for all Monolingual Information Retrieval tasks of the Cross-Language Evaluation Forum (CLEF) 2002. Enabling stemming in SearchServer increased average precision by 16 points in Finnish, 9 points in German, 4 points in Spanish, 3 points in Dutch, 2 points in French and Italian, and 1 point in Swedish and English. Accent-indexing increased average precision by 3 points in Finnish and 2 points in German, but decreased it by 2 points in French and 1 point in Italian and Swedish. Treating apostrophes as word separators increased average precision by 3 points in French and 1 point in Italian. Confidence intervals produced using the bootstrap percentile method were found to be very similar to those produced using the standard method; both were of similar width to rank-based intervals for differences in average precision, but substantially narrower for differences in Precision@10.

1 Introduction

Hummingbird SearchServer[1] is an indexing, search and retrieval engine for embedding in Windows and UNIX information applications. SearchServer, originally a product of Fulcrum Technologies, was acquired by Hummingbird in 1999. Founded in 1983 in Ottawa, Canada, Fulcrum produced the first commercial application program interface (API) for writing information retrieval applications, Fulcrum® Ful/Text™. The SearchServer kernel is embedded in many Hummingbird products, including SearchServer, an application toolkit used for knowledge-intensive applications that require fast access to unstructured information.

SearchServer supports a variation of the Structured Query Language (SQL), SearchSQL™, which has extensions for text retrieval. SearchServer conforms to subsets of the Open Database Connectivity (ODBC) interface for C programming language applications and the Java Database Connectivity (JDBC) interface for Java applications. Almost 200 document formats are supported, such as Word, WordPerfect, Excel, PowerPoint, PDF and HTML.

SearchServer works in Unicode internally [3] and supports most of the world's major character sets and languages. The major conferences in text retrieval evaluation (CLEF [2], NTCIR [5] and TREC [10]) have provided opportunities to objectively evaluate SearchServer's support for a dozen languages. This paper focuses on evaluating SearchServer's options for 8 European languages using the CLEF 2002 test collections.

[1] Fulcrum® is a registered trademark, and SearchServer™, SearchSQL™, Intuitive Searching™ and Ful/Text™ are trademarks of Hummingbird Ltd. All other copyrights, trademarks and tradenames are the property of their respective owners.

Table 1: Sizes of CLEF 2002 Document Sets

Language	Text Size (uncompressed)	Number of Documents
German	555,285,140 bytes (530 MB)	225,371
Spanish	544,347,121 bytes (519 MB)	215,738
Dutch	558,560,087 bytes (533 MB)	190,604
Swedish	374,371,465 bytes (357 MB)	142,819
English	441,048,231 bytes (421 MB)	113,005
Italian	290,771,116 bytes (277 MB)	108,578
French	253,528,734 bytes (242 MB)	87,191
Finnish	143,902,109 bytes (137 MB)	55,344

2 Setup

For the experiments described in this paper, an internal development build of SearchServer 5.3 was used (5.3.500.279).

2.1 Data

The CLEF 2002 document sets consisted of tagged (SGML-formatted) news articles (mostly from 1994) in 8 different languages: German, French, Italian, Spanish, Dutch, Swedish, Finnish and English. Table 1 gives their sizes. For more information on the CLEF collections, see the CLEF web site [2].

2.2 Text Reader

The custom text reader called cTREC, originally written for handling TREC collections [12], handled expansion of the library files of the CLEF collections and was extended to support the CLEF guidelines of only indexing specific fields of specific documents. The entities described in the DTD files were also converted, e.g. "=" was converted to the equal sign "=".

The documents were assumed to be in the Latin-1 character set, the code page which, for example, assigns e-acute (é) hexadecimal 0xe9 or decimal 233. cTREC passes through the Latin-1 characters, i.e. does not convert them to Unicode. SearchServer's Translation Text Reader (nti), was chained on top of cTREC and the Win_1252_UCS2 translation was specified via its /t option to translate from Latin-1 to the Unicode character set desired by SearchServer.

2.3 Indexing

A separate SearchServer table was created for each language, created with a SearchSQL statement such as the following:

```
CREATE SCHEMA CLEF02DE CREATE TABLE CLEF02DE
(DOCNO VARCHAR(256) 128)
TABLE_LANGUAGE 'GERMAN'
STOPFILE 'LANGDE.STP'
PERIODIC
BASEPATH 'e:\data\clef';
```

The TABLE_LANGUAGE parameter specifies which language to use when performing stemming operations at index time. The STOPFILE parameter specifies a stop file containing typically a couple hundred stop words to not index; the stop file also contains instructions on changes to the default indexing rules, for example, to enable accent-indexing, or to change the apostrophe to a word separator. Here are the first few lines of the stop file used for the French task:

```
IAC = "\u0300-\u0345"
PST = "’‘"
STOPLIST =
a
à
afin
```

The IAC line enables indexing of the specified accents (Unicode combining diacritical marks 0x0300-0x0345). Accent-indexing was enabled for all runs except the Italian and English runs. Accents were known to be specified in the Italian queries but were not consistently used in the Italian documents. The PST line adds the specified characters (apostrophes in this case) to the list of word separators. The apostrophes were changed to word separators except for English runs.

Into each table, we just needed to insert one row, specifying the top directory of the library files for the language, using an Insert statement such as the following:

```
INSERT INTO CLEF02DE ( FT_SFNAME, FT_FLIST ) VALUES
('German','cTREC/E/d=128:s!nti/t=Win_1252_UCS2:cTREC/C/@:s');
```

To index each table, we just executed a Validate Index statement such as the following:

```
VALIDATE INDEX CLEF02DE VALIDATE TABLE;
```

By default, the index supports both exact matching (after some Unicode-based normalizations, such as converting to upper-case and decomposed form) and matching on stems.

3 Search Techniques

The CLEF organizers created 50 “topics” (numbered 91-140) and translated them into many languages. Each topic contained a “Title” (subject of the topic), “Description” (a one-sentence specification of the information need) and “Narrative” (more detailed guidelines for what a relevant document should or should not contain). The participants were asked to use the Title and Description fields for at least one automatic submission per task this year to facilitate comparison of results.

We created an ODBC application, called QueryToRankings.c, based on the example stsample.c program included with SearchServer, to parse the CLEF topics files, construct and execute corresponding SearchSQL queries, fetch the top 1000 rows, and write out the rows in the results format requested by CLEF. SELECT statements were issued with the SQLExecDirect api call. Fetches were done with SQLFetch (typically 1000 SQLFetch calls per query).

3.1 Intuitive Searching

For all runs, we used SearchServer’s Intuitive Searching, i.e. the IS_ABOUT predicate of SearchSQL, which accepts unstructured text. For example, for the German version of topic 41 (from last year), the Title was “Pestizide in Babykost” (Pesticides in Baby Food), and the Description was “Berichte über Pestizide in Babynahrung sind gesucht” (Find reports on pesticides in baby food). A corresponding SearchSQL query would be:

```
SELECT RELEVANCE('V2:3') AS REL, DOCNO
FROM CLEF02DE
WHERE FT_TEXT IS_ABOUT 'Pestizide in Babykost Berichte über
Pestizide in Babynahrung sind gesucht'
ORDER BY REL DESC;
```

This query would create a working table with the 2 columns named in the SELECT clause, a

REL column containing the relevance value of the row for the query, and a DOCNO column containing the document's identifier. The ORDER BY clause specifies that the most relevant rows should be listed first. The statement "SET MAX_SEARCH_ROWS 1000" was previously executed so that the working table would contain at most 1000 rows.

3.2 Stemming

SearchServer "stems" each distinct word to one or more base forms, called stems. For example, in English, "baby", "babied", "babies", "baby's" and "babying" all have "baby" as a stem. Compound words in German, Dutch and Finnish produce multiple stems; e.g., in German, "babykost" has "baby" and "kost" as stems. SearchServer 5.3 uses the lexicon-based Inxight LinguistX Platform 3.3.1 for stemming operations.

By default, Intuitive Searching stems each word in the query, counts the number of occurrences of each stem, and creates a vector. Optionally some stems are discarded (secondary term selection) if they have a high document frequency or to enforce a maximum number of stems, but we didn't discard any for our CLEF runs. The index is searched for documents containing terms which stem to any of the stems of the vector.

The VECTOR_GENERATOR set option controls which stemming operations are performed by Intuitive Searching. To enable stemming, we used the same setting for each language except for the /lang parameter. For example, for German, the setting was 'word!ftelp/lang=german/base/noalt | * | word!ftelp/lang=german/inflect'. To disable stemming, the setting was just ''.

Besides linguistic expansion from stemming, we did not do any other kinds of query expansion. For example, we did not use approximate text searching for spell-correction because the queries were believed to be spelled correctly. We did not use row expansion or any other kind of blind feedback technique.

3.3 Statistical Relevance Ranking

SearchServer calculates a relevance value for a row of a table with respect to a vector of stems based on several statistics. The inverse document frequency of the stem is estimated from information in the dictionary. The term frequency (number of occurrences of the stem in the row (including any term that stems to it)) is determined from the reference file. The length of the row (based on the number of indexed characters in all columns of the row, which is typically dominated by the external document), is optionally incorporated. The already-mentioned count of the stem in the vector is also used. To synthesize this information into a relevance value, SearchServer dampens the term frequency and adjusts for document length in a manner similar to Okapi [6] and dampens the inverse document frequency in a manner similar to [8]. SearchServer's relevance values are always an integer in the range 0 to 1000.

SearchServer's RELEVANCE_METHOD setting can be used to optionally square the importance of the inverse document frequency (by choosing a RELEVANCE_METHOD of 'V2:4' instead of 'V2:3'). The importance of document length to the ranking is controlled by SearchServer's RELEVANCE_DLEN_IMP setting (scale of 0 to 1000). For all runs in this paper, RELEVANCE_METHOD was set to 'V2:3' and RELEVANCE_DLEN_IMP was set to 750.

3.4 Query Stop Words

Our QueryToRankings program removed words such as "find", "relevant" and "document" from the topics before presenting them to SearchServer, i.e. words which are not stop words in general but were commonly used in the CLEF topics as general instructions. For the submitted runs, the lists were developed by examining the CLEF 2000 and 2001 topics (not this year's topics). For the diagnostic runs in this paper, "finde" was added as a query stop word because it was noticed to be common in the German topics this year. An evaluation of the impact of query stop words is provided below.

Table 2: Precision with Stemming Enabled and Disabled

Run	AvgP	P@5	P@10	P@20	Rec0	Rec30	#Topics
Finnish	0.393	44.0%	36.0%	28.2%	0.707	0.502	30
	0.232	29.3%	23.3%	17.7%	0.540	0.299	
German	0.442	64.4%	55.4%	46.7%	0.819	0.538	50
	0.348	55.2%	48.2%	38.7%	0.726	0.426	
Spanish	0.491	68.0%	58.8%	51.2%	0.871	0.617	50
	0.454	64.8%	57.2%	50.9%	0.833	0.586	
Dutch	0.442	58.4%	50.6%	41.3%	0.822	0.529	50
	0.410	54.8%	48.4%	40.6%	0.779	0.516	
French	0.428	52.8%	44.6%	35.6%	0.774	0.554	50
	0.404	49.6%	39.8%	32.6%	0.824	0.488	
Italian	0.409	50.2%	45.3%	36.0%	0.740	0.537	49
	0.395	51.4%	43.3%	35.4%	0.760	0.491	
Swedish	0.348	44.1%	37.6%	29.4%	0.754	0.439	49
	0.334	44.5%	37.1%	28.8%	0.705	0.436	
English	0.508	57.6%	45.2%	34.3%	0.909	0.682	42
	0.500	55.2%	42.9%	33.2%	0.894	0.644	

4 Results

The evaluation measures are likely explained in an appendix of this volume. Briefly: "Precision" is the percentage of retrieved documents which are relevant. "Precision@n" is the precision after n documents have been retrieved. "Average precision" for a topic is the average of the precision after each relevant document is retrieved (using zero as the precision for relevant documents which are not retrieved). "Recall" is the percentage of relevant documents which have been retrieved. "Interpolated precision" at a particular recall level for a topic is the maximum precision achieved for the topic at that or any higher recall level. For a set of topics, the measure is the average of the measure for each topic (i.e. all topics are weighted equally).

The Monolingual Information Retrieval tasks were to run 50 queries against document collections in the same language and submit a list of the top-1000 ranked documents to CLEF for judging (in June 2002). CLEF produced a "qrels" file for each of the 8 tasks: a list of documents judged to be relevant or not relevant for each topic. From these, the evaluation measures were calculated with Chris Buckley's trec_eval program.

For some topics and languages, no documents were judged relevant. The precision scores are just averaged over the number of topics for which at least one document was judged relevant.

4.1 Impact of Stemming

Table 2 shows two runs for each language. The first run uses the same settings as were used for the submitted runs which used the Title and Description fields; in particular, stemming was enabled. The second run uses the same settings except that VECTOR_GENERATOR was set to the empty string, which disables stemming. Listed for each run are its average precision (AvgP), the precision after 5, 10 and 20 documents retrieved (P@5, P@10 and P@20 respectively), and the interpolated precision at 0% and 30% recall (Rec0 and Rec30 respectively). Additionally listed for the runs with stemming enabled is the number of topics which contained at least one relevant document for that language. The languages are ordered by descending difference in average precision. Stemming increased average precision in Finnish by 69%, German 27%, Spanish 8%, Dutch 8%, French 6%, Italian 4%, Swedish 4% and English 2%.

Table 3: Impact of Stemming on Average Precision

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
Finnish-td	0.161	(0.097, 0.232)	24-5-1	0.753 (115), 0.453 (122)
German-td	0.093	(0.052, 0.136)	35-15-0	0.553 (105), 0.410 (140)
Spanish-td	0.038	(0.012, 0.064)	32-18-0	0.337 (119), 0.248 (137)
Dutch-td	0.032	(0.001, 0.065)	31-17-2	0.437 (116), 0.350 (109)
French-td	0.024	(−0.007, 0.059)	23-24-3	0.406 (115), 0.322 (140)
Italian-td	0.015	(−0.006, 0.037)	27-22-0	0.185 (140), 0.178 (137)
Swedish-td	0.014	(−0.004, 0.032)	27-18-4	−0.217 (93), 0.204 (129)
English-td	0.008	(−0.020, 0.035)	22-16-4	−0.283 (139), 0.213 (129)

Table 4: Impact of Stemming on Average Precision, Title-only queries

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
Finnish-t	0.175	(0.103, 0.259)	26-4-0	0.799 (115), 0.797 (98)
German-t	0.108	(0.057, 0.167)	36-13-1	1.000 (137), 0.521 (140)
Spanish-t	0.036	(0.016, 0.061)	30-8-12	0.485 (119), 0.201 (139)
Italian-t	0.035	(0.012, 0.062)	20-17-12	0.377 (115), 0.263 (137)
Swedish-t	0.033	(0.013, 0.055)	25-7-17	0.323 (129), 0.250 (134)
Dutch-t	0.030	(0.000, 0.065)	26-19-5	0.567 (116), −0.249 (137)
French-t	0.030	(0.005, 0.060)	24-19-7	0.506 (115), 0.261 (103)
English-t	−0.002	(−0.028, 0.025)	20-14-8	0.312 (103), −0.276 (139)

Most of the remaining tables will focus on one particular precision measure (usually average precision), comparing the scores when a particular feature (such as stemming) is enabled to when it is disabled. The columns of these tables are as follows:

- "Experiment" is the language and topic fields used (for example, "-td" indicates the Title and Description fields were used).
- "AvgDiff" is the average difference in the precision score. In [9], a difference of at least 2 full points (i.e. >=0.020) is considered "noticeable", 4 points "material", 6 points "striking" and 8 points "dramatic".
- "95% Confidence" is an approximate 95% confidence interval for the average difference calculated using the bootstrap percentile method (described in the last section). 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, 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 precision was higher, lower and tied (respectively) with the feature enabled. These numbers should always add to the number of topics for the language (as per Table 2).
- "2 Largest Diffs (Topic)" lists the two largest differences in the precision score (based on the absolute value), with each followed by the corresponding topic number in brackets (the topic numbers range from 91 to 140).

Table 3 shows the impact of stemming on the average precision measure. The benefit for Finnish and German, for which stemming includes compound-breaking, is dramatic. For example, Finnish topic 115, regarding "avioerotilastoja" (divorce statistics), apparently benefits from compound-breaking. Surprisingly, the other investigated language for which compounds are broken, Dutch,

Table 5: Impact of Discarding Query Stop Words on Average Precision

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
Spanish-td	0.016	(0.007, 0.027)	24-4-22	0.151 (138), 0.137 (100)
English-td	0.014	(−0.001, 0.032)	17-8-17	0.250 (137), 0.149 (106)
Finnish-td	0.009	(0.002, 0.020)	11-1-18	0.132 (122), 0.050 (114)
Italian-td	0.007	(−0.002, 0.016)	20-8-21	0.113 (93), −0.107 (91)
German-td	0.006	(0.000, 0.014)	19-15-16	0.099 (138), 0.091 (99)
Swedish-td	0.006	(−0.001, 0.015)	17-10-22	0.106 (111), 0.100 (132)
Dutch-td	0.001	(−0.003, 0.006)	15-12-23	0.062 (138), −0.049 (123)
French-td	−0.000	(−0.013, 0.012)	18-10-22	−0.167 (123), −0.150 (132)

does not similarly stand out, unlike last year [11], though its confidence interval still overlaps the one for German.

Table 4 shows the impact of stemming on the shorter (Title-only) queries. It appears the benefits are a little bigger for the shorter queries in most languages, with English the only language without a noticeable benefit on average. Of course, stemming can hurt precision for some queries, as in English topic 139 (EU fishing quotas), so an application probably should make stemming a user-controllable option.

4.2 Impact of Query Stop Words

Table 5 shows the impact of discarding query stop words, such as "find", "relevant" and "documents". Query stop words differ from general stop words (such as "the", "of", "by") in that they do not seem to be noise words in general, but their common use in past CLEF topic sets (particularly the Description and Narrative fields) suggests they are likely not useful terms when encountered in CLEF queries. In the table, a positive difference indicates a benefit from removing query stop words from the topics.

Table 5 shows that the impact of discarding query stop words was always minor (the biggest average benefit was just 1.6 points), though some of the differences are "statistically significant" because of the consistency of the minor benefits. This is a case where a "statistically significant" benefit is still not a "significant" benefit.

Sometimes noise words may occur in relevant documents by chance and scores may fall if the noise words are discarded. Apparently that happened in French topics 123 and 132 (regarding "mariage" and "Kaliningrad" respectively) in which excluding "trouver" and "documents" decreased the scores, even though they don't seem to be meaningful terms for their queries.

4.3 Impact of Stop Words

Tables 6 and 7 show the impact of using stop words on the average precision measure. To do this experiment, two tables were created for each language, one indexed with a stopfile containing typically a couple hundred stop words, the other with no stop words (though other SearchServer stopfile instructions, such as accent-indexing and apostrophes as word separators, were kept the same as used for the submitted runs). For this experiment, query stop words were not discarded for either run, to isolate the impact of the general stop words on precision. In the tables, a positive difference indicates a benefit to specifying stop words.

Table 6 shows the impact of using stop words for Title plus Description queries was very slight on average, and none of the differences were statistically significant. Table 7 shows there was a noticeable benefit for full topic queries (i.e. when additionally including the Narrative) for some languages, and a statistically significant benefit for most of them. Other benefits of specifying stop words are to reduce search time, indexing time and index size. However, there may be cases when what is usually a stop word is meaningful to a query (e.g. find documents containing "to be

Table 6: Impact of Stop Words on Average Precision

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
Spanish-td	0.006	(−0.004, 0.018)	25-24-1	0.186 (99), 0.125 (98)
French-td	0.005	(−0.002, 0.015)	23-22-5	0.167 (123), 0.099 (105)
German-td	0.004	(−0.001, 0.011)	26-21-3	0.064 (106), 0.050 (99)
Finnish-td	0.004	(−0.007, 0.015)	17-9-4	−0.088 (139), 0.083 (132)
Swedish-td	0.002	(−0.003, 0.009)	20-23-6	0.071 (99), −0.054 (130)
Dutch-td	−0.000	(−0.010, 0.009)	28-19-3	−0.117 (138), −0.114 (104)
Italian-td	−0.000	(−0.011, 0.010)	27-22-0	−0.164 (96), −0.083 (139)
English-td	−0.001	(−0.011, 0.011)	14-22-6	0.167 (126), −0.114 (133)

Table 7: Impact of Stop Words on Average Precision, Full Topic Queries

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
Spanish-tdn	0.023	(0.012, 0.034)	36-13-1	0.156 (99), 0.126 (98)
Italian-tdn	0.020	(0.007, 0.034)	33-16-0	0.172 (132), −0.127 (96)
Swedish-tdn	0.017	(0.008, 0.028)	33-14-2	0.163 (102), 0.096 (96)
French-tdn	0.016	(0.004, 0.033)	34-12-4	0.303 (109), 0.094 (128)
German-tdn	0.015	(0.003, 0.029)	33-14-3	0.262 (137), 0.080 (102)
Finnish-tdn	0.012	(0.004, 0.020)	20-7-3	0.083 (132), 0.054 (116)
English-tdn	0.007	(−0.001, 0.018)	20-15-7	0.164 (126), 0.101 (99)
Dutch-tdn	0.005	(−0.013, 0.023)	29-17-4	0.213 (116), −0.177 (109)

or not to be"), so it may be better to make stop word elimination an option at search time rather than at index time, depending on the goals of the application.

Stop word lists for many languages are on the Neuchâtel resource page [7]. Our stop word lists may contain differences.

4.4 Impact of Indexing Accents

Tables 8 and 9 show the impact of accent-indexing on the average precision measure. To do this experiment, two tables were created for each language, one preserving accents (e.g. "bébé" and "bebe" would be distinct words) and one which dropped the accents (e.g. "bébé" would be the same as "bebe"). Of course, at search-time SearchServer uses the same rules as at index-time. For this experiment, no stop words were used and no query stop words were discarded. Otherwise, the

Table 8: Impact of Preserving Accents on Average Precision

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
Finnish-td	0.026	(−0.003, 0.060)	17-11-2	0.289 (109), 0.255 (139)
German-td	0.018	(−0.002, 0.040)	29-19-2	0.267 (108), 0.256 (135)
Dutch-td	0.002	(−0.001, 0.005)	12-6-32	0.050 (110), 0.031 (127)
English-td	0.000	(−0.001, 0.001)	0-0-42	0.000 (116), 0.000 (92)
Spanish-td	−0.003	(−0.034, 0.024)	23-25-2	−0.572 (98), 0.298 (103)
Swedish-td	−0.013	(−0.028, 0.000)	15-26-8	−0.167 (134), −0.160 (127)
Italian-td	−0.014	(−0.038, 0.000)	10-25-14	−0.502 (98), −0.109 (101)
French-td	−0.018	(−0.049, 0.009)	16-28-6	−0.474 (98), −0.297 (94)

Table 9: Impact of Preserving Accents on Average Precision, Title-only queries

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
Finnish-t	0.061	(0.008, 0.128)	14-14-2	0.703 (98), 0.540 (139)
German-t	0.007	(−0.048, 0.065)	22-16-12	1.000 (137), −0.873 (127)
Dutch-t	0.000	(−0.001, 0.001)	5-2-43	0.007 (127), 0.002 (103)
English-t	0.000	(−0.001, 0.001)	0-0-42	0.000 (116), 0.000 (92)
Spanish-t	−0.002	(−0.045, 0.032)	20-15-15	−0.799 (98), 0.477 (103)
Swedish-t	−0.008	(−0.037, 0.017)	8-10-31	−0.496 (127), 0.329 (129)
Italian-t	−0.014	(−0.041, 0.001)	5-5-39	−0.666 (98), 0.015 (128)
French-t	−0.024	(−0.065, 0.008)	16-19-15	−0.709 (98), −0.462 (94)

Table 10: Impact of Dropping Apostrophes on Average Precision

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
French-td	0.034	(0.007, 0.075)	28-17-5	0.833 (121), 0.220 (105)
Italian-td	0.011	(−0.001, 0.026)	23-18-8	0.211 (93), 0.195 (139)
Dutch-td	0.003	(−0.003, 0.012)	7-8-35	0.188 (98), −0.048 (130)
English-td	0.002	(−0.001, 0.005)	13-18-11	0.039 (103), 0.038 (135)
Spanish-td	0.000	(−0.001, 0.001)	1-4-45	0.004 (101), −0.001 (91)
Swedish-td	0.000	(−0.001, 0.001)	2-0-47	0.004 (130), 0.001 (139)
German-td	−0.000	(−0.001, 0.001)	3-3-44	−0.004 (130), 0.002 (91)
Finnish-td	−0.000	(−0.001, 0.001)	0-3-27	−0.001 (124), −0.000 (92)

settings were the same as for the submitted runs; in particular, apostrophes were used as word separators except in English.

Tables 8 and 9 show that topic 98, regarding the Kaurismäki brothers, was strongly affected in many languages by whether or not accents were preserved. Spanish, French and Italian topics 98 included the accent in Kaurismäki, but the documents more often did not include the accent, so accent-indexing hurt precision in those cases. But accent-indexing was helpful for Finnish for this topic, apparently because in Finnish there were variants which required stemming to match (e.g. Kaurismäkien and Kaurismäen), and the stemmer was more effective when given the words with the accents preserved. It appears it would help if the stemmer was modified to be more tolerant of missing accents.

4.5 Impact of Apostrophes as Word Separators

Table 10 shows the impact of treating apostrophes as word separators on the average precision measure. To do this experiment, two tables were created for each language, one treating apostrophes as word separators, the other not. No stop words were used and no query stop words were discarded. Otherwise, the settings were the same as for the submitted runs; in particular, accent-indexing was enabled except in Italian and English.

Table 10 shows that treating apostrophes as word separators had a noticeable benefit for French. For example, French topic 121 may be benefiting from breaking “d’Ayrton” at its apostrophe. The benefit for Italian may have been less because stemming appears to be handling apostrophes. For example, in Italian, if apostrophes are not word separators, “l’ombrello” still matches “ombrello” when stemming is enabled, whereas in French, “l’école” still does not match “école” (again, this difference is moot when apostrophes are treated as word separators). The impact for other languages is slight, including for English.

Table 11: Comparison of Submitted Runs with Medians in Average Precision

Submission	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
Swedish-tdn	0.125	(0.082, 0.173)	37-6-6	0.762 (101), 0.500 (134)
Swedish-td	0.093	(0.059, 0.132)	37-3-9	0.589 (101), 0.433 (129)
Dutch-tdn	0.080	(0.050, 0.110)	37-7-6	0.366 (138), 0.261 (117)
Finnish-td	0.076	(0.022, 0.138)	20-7-3	0.641 (115), 0.364 (122)
Finnish-tdn	0.068	(0.006, 0.136)	20-8-2	0.618 (126), 0.485 (115)
Spanish-td	0.042	(0.012, 0.070)	37-12-1	−0.326 (138), 0.262 (120)
Dutch-td	0.039	(0.018, 0.059)	35-10-5	0.213 (111), −0.212 (102)
French-td	0.017	(−0.010, 0.049)	24-17-9	0.500 (121), −0.274 (94)
German-td	0.013	(−0.013, 0.039)	26-20-4	0.297 (105), 0.244 (137)
Italian-td	−0.001	(−0.025, 0.025)	20-25-4	0.254 (98), 0.240 (137)

4.6 Submitted runs

We submitted 10 monolingual runs (the maximum allowed) in June 2002. Runs humDE02, humFR02, humIT02, humES02, humNL02, humSV02 and humFI02 provided a run for each language using the Title and Description fields as requested by the organizers (note that English monolingual runs were not accepted). For the remaining 3 runs, we submitted an extra run for Finnish, Swedish and Dutch including the Narrative field (runs humFI02n, humSV02n, humNL02n); these languages were expected to have the fewest participants, so additional submissions seemed more likely to be helpful for the judging pools. The precision scores of the submitted runs are expected to be included in an appendix of this volume. Table 11 shows a comparison of the submitted runs with the median scores of submitted monolingual runs for each language. In all but one case, SearchServer scored higher than the median on more topics than it scored lower. Note that the relative performance on different languages may not be meaningful for several reasons, including that the medians are from a mix of runs where some may have used the Narrative field, multiple runs may be submitted by the same group, and the mixture can vary across languages.

The submitted runs of June used an older, experimental build than was used for the diagnostic runs in August, and there may be minor differences in the scores even when the settings are the same.

5 Confidence Intervals for Precision Differences

The 95% confidence intervals presented in this paper have been produced using Efron's Bootstrap (percentile method). If there are 50 topics (i.e. 50 precision differences), then precision differences are chosen randomly (with replacement) 50 times, producing a "bootstrap sample", and a mean (average) is computed from this sample. This step is repeated B times (e.g. B=100,000). The B sample means are sorted, the bottom and top 2.5% are discarded, and the endpoints of the remaining range of sample means are an approximate 95% confidence interval for the average difference in precision (we always rounded so that the listed endpoints are not actually in the produced interval). The bootstrap percentile method is considered to work well in more cases than the standard method of using the mean plus/minus 1.96 times the standard error, though there are more complicated bootstrap methods which are considered even more general [1].

Table 12 shows the bootstrap confidence intervals produced for the impact of stemming on average precision with different numbers of iterations. Even at just 1000 iterations the values are fairly close to the values at 1 million iterations. When comparing 1,000,000 iterations to 100,000, very few of the endpoints changed, and they only changed by 0.001. For the confidence intervals in this paper, we used B=100,000.

Tables 13 and 14 contain side-by-side comparisons of the approximate 95% confidence intervals

Table 12: Impact of Number of Bootstrap Iterations on Confidence Intervals

Experiment	B=1000	B=10,000	B=100,000	B=1,000,000
Finnish-td	(0.102, 0.232)	(0.099, 0.232)	(0.097, 0.232)	(0.097, 0.231)
German-td	(0.054, 0.136)	(0.053, 0.137)	(0.052, 0.136)	(0.052, 0.137)
Spanish-td	(0.012, 0.066)	(0.012, 0.064)	(0.012, 0.064)	(0.012, 0.065)
Dutch-td	(0.002, 0.060)	(0.001, 0.065)	(0.001, 0.065)	(0.001, 0.065)
French-td	(−0.006, 0.058)	(−0.007, 0.059)	(−0.007, 0.059)	(−0.007, 0.058)
Italian-td	(−0.005, 0.036)	(−0.006, 0.037)	(−0.006, 0.037)	(−0.006, 0.037)
Swedish-td	(−0.005, 0.032)	(−0.004, 0.032)	(−0.004, 0.032)	(−0.004, 0.032)
English-td	(−0.021, 0.035)	(−0.020, 0.036)	(−0.020, 0.035)	(−0.020, 0.035)

Table 13: Comparison of Confidence Intervals for Impact of Stemming on Average Precision

Experiment	AvgDiff	Bootstrap 95% Confidence	+/− (1.96 * StdErr)	Wilcoxon EstDiff	95% Confidence
Finnish-td	0.161	(0.097, 0.232)	(0.093, 0.229)	0.150	(0.073, 0.215)
German-td	0.093	(0.052, 0.136)	(0.050, 0.136)	0.077	(0.032, 0.133)
Spanish-td	0.038	(0.012, 0.064)	(0.011, 0.064)	0.028	(0.008, 0.053)
Dutch-td	0.032	(0.001, 0.065)	(−0.001, 0.064)	0.017	(0.000, 0.040)
French-td	0.024	(−0.007, 0.059)	(−0.009, 0.057)	0.008	(−0.009, 0.042)
Italian-td	0.015	(−0.006, 0.037)	(−0.007, 0.037)	0.010	(−0.007, 0.034)
Swedish-td	0.014	(−0.004, 0.032)	(−0.004, 0.032)	0.009	(−0.001, 0.026)
English-td	0.008	(−0.020, 0.035)	(−0.019, 0.036)	0.005	(−0.013, 0.027)

Table 14: Comparison of Confidence Intervals for Impact of Stemming on Precision@10

Experiment	AvgDiff	Bootstrap 95% Confidence	+/− (1.96 * StdErr)	Wilcoxon EstDiff	95% Confidence
Finnish-td	0.127	(0.046, 0.211)	(0.044, 0.210)	0.100	(0.000, 0.250)
German-td	0.072	(0.013, 0.133)	(0.011, 0.133)	0.050	(0.000, 0.150)
French-td	0.048	(0.015, 0.083)	(0.015, 0.081)	0.050	(0.000, 0.100)
English-td	0.024	(−0.005, 0.055)	(−0.006, 0.054)	0.000	(−0.050, 0.050)
Dutch-td	0.022	(−0.016, 0.063)	(−0.017, 0.062)	0.000	(−0.050, 0.100)
Italian-td	0.020	(−0.004, 0.045)	(−0.005, 0.046)	0.000	(−0.050, 0.050)
Spanish-td	0.016	(−0.012, 0.045)	(−0.012, 0.044)	0.000	(−0.050, 0.050)
Swedish-td	0.004	(−0.033, 0.041)	(−0.032, 0.041)	0.000	(−0.050, 0.050)

produced by the bootstrap percentile method and the standard method. It turns they are very similar. There is a disagreement on statistical significance (i.e. when zero is not in the interval) in the case of Dutch in Table 13, but it is a borderline case.

Tables 13 and 14 also include an estimator and 95% confidence interval based on the Wilcoxon signed rank test (the 2 rightmost columns). (We implemented an exact computation, including for the case of ties in the absolute values of the differences [4]). For differences in average precision, the widths of the intervals are very similar (the bootstrap intervals are a little smaller than the Wilcoxon intervals for the Finnish and German results, and the Wilcoxon intervals are a little smaller for the others); the methods agree on which differences are statistically significant. However, for differences in Precision@10, the bootstrap intervals are a lot smaller than the Wilcoxon intervals (because the Wilcoxon is based on ranks, it cannot distinguish between a shift of 0.01 and 0.09 (they have the same effect on the ranks because every difference is a multiple of 0.10)); the methods still agree on statistically significant results (for the 8 cases listed).

References

[1] Michael R. Chernick. Bootstrap Methods: A Practitioner's Guide. 1999. John Wiley & Sons.

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

[3] Andrew Hodgson. Converting the Fulcrum Search Engine to Unicode. In Sixteenth International Unicode Conference, Amsterdam, The Netherlands, March 2000.

[4] Myles Hollander and Douglas A. Wolfe. Nonparametric Statistical Methods. Second Edition, 1999. John Wiley & Sons.

[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, M. Gatford. (City University.) Okapi at TREC-3. In D. K. Harman, editor, Overview of the Third Text REtrieval Conference (TREC-3). NIST Special Publication 500-226. http://trec.nist.gov/pubs/trec3/t3_proceedings.html

[7] Jacques Savoy. (Université de Neuchâtel.) CLEF and Multilingual information retrieval resource page. http://www.unine.ch/info/clef/

[8] Amit Singhal, John Choi, Donald Hindle, David Lewis and Fernando Pereira. AT&T at TREC-7. In E. M. Voorhees and D. K. Harman, editors, Proceedings of the Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242. http://trec.nist.gov/pubs/trec7/t7_proceedings.html

[9] K. Sparck Jones, S. Walker and S. E. Robertson. (City University.) A probabilistic model of information retrieval: development and status. August 1998. Page 15.

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

[11] Stephen Tomlinson. Stemming Evaluated in 6 Languages by Hummingbird SearchServer™ at CLEF 2001. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck, editors, Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001. Revised Papers. Springer LNCS 2406. http://link.springer.de/link/service/series/0558/tocs/t2406.htm

[12] Stephen Tomlinson and Tom Blackwell. Hummingbird's Fulcrum SearchServer at TREC-9. In E. M. Voorhees and D. K. Harman, editors, Proceedings of the Ninth Text REtrieval Conference (TREC-9). NIST Special Publication 500-249. http://trec.nist.gov/pubs/trec9/t9_proceedings.html