European Web Retrieval Experiments at WebCLEF 2006

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

August 20, 2006

Abstract

Hummingbird participated in the WebCLEF mixed monolingual retrieval task of the Cross-Language Evaluation Forum (CLEF) 2006. In this task, the system was given 1939 known-item queries, and the goal was to find the desired page in the 82GB EuroGOV collection (3.4 million pages crawled from government sites of 27 European domains). The 1939 queries included 124 new manually-created queries, 195 manually-created queries from last year, and 1620 automatically-generated queries. In our experiments, the results on the automatically-generated queries were not always predictive of the results on the manually-created queries; in particular, our title-weighting and duplicate-filtering techniques were fairly effective on the manually-created queries but were detrimental on the automatically-generated queries.

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

First Relevant Score, Automatically-Generated Queries

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 [10]) have provided judged test collections for objective experimentation with SearchServer in more than a dozen languages.

This (draft) paper describes experimental work with SearchServer for the task of finding named pages in various European languages using the WebCLEF 2006 test collection.

[1] SearchServer™, SearchSQL™and Intuitive Searching™ are trademarks of Hummingbird Ltd. All other copyrights, trademarks and tradenames are the property of their respective owners.

2 Methodology

For the submitted runs in July 2006, SearchServer experimental development build 7.0.1.271 was used.

2.1 Data

The collection to be searched was the EuroGOV collection [8]. It consisted of 3,589,502 pages crawled from government sites of 27 European domains. Uncompressed, it was 88,062,007,676 bytes (82.0 GB). The average document size was 24,533 bytes. Note that we only indexed 3,417,463 of the pages because the organizers provided a "blacklist" of 172,039 pages to omit (primarily binary documents).

For the mixed monolingual task, there were 1939 queries, including 124 new manually-created queries, 195 manually-created queries from last year, and 1620 automatically-generated queries.

Based on the official query labels, here is the count of the number of queries of each language:

- "manual new" topics (124 total):
  - DE 30, EN 30, ES 24, HU 10, NL 30.
- "manual old" topics (195 total):
  - DA 15, DE 30, EN 30, ES 30, HU 15, NL 30, PT 30, RU 15.
- "auto uni" topics (810 total):
  - CS 1, DA 23, DE 61, EL 14, EN 94, ES 13, ET 24, FI 27, FR 51, GA 2, HU 28, IT 12, LT 5, LV 21, NL 14, PL 22, PT 22, RU 3, SK 13, SV 19, UNKNOWN 341.
- "auto bi" topics (810 total):
  - DA 19, DE 61, EL 17, EN 101, ES 19, ET 16, FI 21, FR 60, GA 2, HU 30, IS 1, IT 15, LT 11, LV 24, NL 16, PL 19, PT 26, RU 3, SK 13, SV 21, UNKNOWN 315.

More details on the mixed monolingual task are presumably in the track overview paper.

3 Indexing

Our indexing approach was similar to what we used last year (described in detail in [12]). Briefly, in addition to full-text indexing, the custom text reader cTREC populated particular columns such as TITLE (if any), URL, URL_TYPE and URL_DEPTH. The URL_TYPE was set to ROOT, SUBROOT, PATH or FILE, based on the convention which worked well in TREC 2001 for the Twente/TNO group [16] on the entry page finding task (also known as the home page finding task). The URL_DEPTH was set to a term indicating the depth of the page in the site. Table 1 contains URL types and depths for example URLs. The exact rules we used are given in [13].

We used the first recognized 'charset' specification in the page (e.g. from the meta http-equiv tag) to indicate from which character set to convert the page to Unicode (Win_1252 was assumed if no charset was specified).

One change from last year was the use of a new stopword list which concatenated stopword lists of 15 European languages (DA, DE, EL, EN, ES, FI, FR, HU, IT, NL, NO, PT, RU, SV, TR).

The apostrophe was treated as a term separator. No accents were indexed. Stemming was not used for any of our runs this year.

Table 1: Examples of URL Type and Depth Values

URL	Type	Depth	Depth Term
http://nasa.gov/	ROOT	1	URLDEPTHA
http://www.nasa.gov/	ROOT	1	URLDEPTHA
http://jpl.nasa.gov/	ROOT	2	URLDEPTHAB
http://fred.jpl.nasa.gov/	ROOT	3	URLDEPTHABC
http://nasa.gov/jpl/	SUBROOT	2	URLDEPTHAB
http://nasa.gov/jpl/fred/	PATH	3	URLDEPTHABC
http://nasa.gov/index.html	ROOT	1	URLDEPTHA
http://nasa.gov/fred.html	FILE	2	URLDEPTHAB

3.1 Searching

We executed 6 runs in July 2006, though only 5 were allowed to be submitted. All 6 are described here.

humWC06nos: This run was the same as humWC06 (described below) except that no stopword list was used. (This run was not submitted.)

humWC06: This submitted run was a plain content search of the baseline table. It used the '2:3' relevance method and document length normalization (SET RELEVANCE_DLEN_IMP 500). Below is an example SearchSQL query:

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM EGOV
WHERE
 (FT_TEXT IS_ABOUT 'Giuseppe Medici')
ORDER BY REL DESC;
```

humWC06p run: This submitted run was the same as humWC06 except that it put additional weight on matches in the title, url, first heading and some meta tags, including extra weight on matching the query as a phrase in these fields. Below is an example SearchSQL query. The searches on the ALL_PROPS column (which contained a copy of the title, url, etc. as described in [13]) are the difference from the humWC06 run. Note that the FT_TEXT column indexed the content and also all of the non-content fields except for the URL. Unlike last year, we used WEIGHT 2 instead of WEIGHT 1 for the "ALL_PROPS IS_ABOUT" weight:

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM EGOV
WHERE
 (ALL_PROPS CONTAINS 'Giuseppe Medici' WEIGHT 1) OR
 (ALL_PROPS IS_ABOUT 'Giuseppe Medici' WEIGHT 2) OR
 (FT_TEXT IS_ABOUT 'Giuseppe Medici' WEIGHT 10)
ORDER BY REL DESC;
```

humWC06dp run: This submitted run was the same as humWC06p except that it put additional weight on urls of depth 4 or less. Less deep urls also received higher weight from inverse document frequency because (presumably) they were less common. Below is an example WHERE clause:

```
WHERE
((ALL_PROPS CONTAINS 'Giuseppe Medici' WEIGHT 1) OR
 (ALL_PROPS IS_ABOUT 'Giuseppe Medici' WEIGHT 2) OR
 (FT_TEXT IS_ABOUT 'Giuseppe Medici' WEIGHT 10)
) AND (
 (URL_TYPE CONTAINS 'ROOT' WEIGHT 0) OR
 (URL_TYPE CONTAINS 'SUBROOT' WEIGHT 0) OR
 (URL_TYPE CONTAINS 'PATH' WEIGHT 0) OR
 (URL_TYPE CONTAINS 'FILE' WEIGHT 0) OR
 (URL_DEPTH CONTAINS 'URLDEPTHA' WEIGHT 5) OR
 (URL_DEPTH CONTAINS 'URLDEPTHAB' WEIGHT 5) OR
 (URL_DEPTH CONTAINS 'URLDEPTHABC' WEIGHT 5) OR
 (URL_DEPTH CONTAINS 'URLDEPTHABCD' WEIGHT 5) )
```

humWC06dpc run: This submitted run was the same as humWC06dp except that it applied an experimental duplicate-filtering heuristic.

humWC06dpcD run: This run was the same as humWC06dpc except that the domain information of the topic metadata was used to restrict the search to the specified domain. Below is an example of the domain filter added to the WHERE clause for a case in which the page was known to be in the 'it' domain (which implied the DOCNO would contain 'Eit').

```
AND (DOCNO CONTAINS 'Eit' WEIGHT 0)
```

4 Results of Web Search Experiments

The 6 runs allow us to isolate 5 'web techniques' which are denoted as follows:

- 's' (stopwords): The humWC06 score minus the humWC06nos score.
- 'p' (extra weight for phrases in the Title and other properties plus extra weight for vector search on properties): The humWC06p score minus the humWC06 score.
- 'd' (modest extra weight for less deep urls): The humWC06dp score minus the humWC06p score.
- 'c' (duplicate-filtering): The humWC06dpc score minus the humWC06dp score.
- 'D' (domain filtering): The humWC06dpcD score minus the humWC06dpc score.

Table 2 lists the mean scores of the 5 submitted runs (and the 1 other diagnostic run in brackets) over the 4 categories of topics:

- "new": the 124 new manually-created topics
- "old": the 195 manually-created topics from last year
- "uni": the 810 automatically-generated "auto uni" topics
- "bi": the 810 automatically-generated "auto bi" topics.

Table 3 isolates the differences in Generalized Success@10 (GS10) between the runs of Table 2. (Details of the column headings can be found in our companion ad hoc paper [11].) 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. Last year [12], GS10 was known as "First Relevant Score" (FRS).

Preliminary findings from Table 3 include the following:

- The 's' technique (stopwords) was not as beneficial on the new topics as last year's topics. We have not yet had time to investigate particular topics to find out why not.
- The 'p' technique (extra weight for phrases in the Title and other properties plus extra weight for vector search on properties), which has been reliably effective on the manually-created queries over the years, was detrimental on the automatically-generated queries.

Table 2: Mean Scores of Submitted WebCLEF Runs

Run	GS10	S1	S5	S10	S50	MRR
humWC06dpcD	0.685	50/124	79/124	88/124	106/124	0.510
humWC06dpc	0.657	49/124	75/124	84/124	102/124	0.494
humWC06dp	0.665	49/124	78/124	86/124	105/124	0.497
humWC06p	0.666	49/124	78/124	86/124	104/124	0.499
humWC06	0.648	44/124	73/124	84/124	101/124	0.466
(humWC06nos)	0.655	43/124	74/124	86/124	102/124	0.463
old topics:						
humWC06dpcD	0.622	72/195	111/195	126/195	158/195	0.463
humWC06dpc	0.610	71/195	110/195	122/195	154/195	0.455
humWC06dp	0.600	71/195	107/195	119/195	159/195	0.447
humWC06p	0.571	67/195	101/195	115/195	154/195	0.425
humWC06	0.528	59/195	94/195	104/195	143/195	0.390
(humWC06nos)	0.524	57/195	93/195	103/195	144/195	0.377
“auto uni” topics:						
humWC06dpcD	0.123	37/810	81/810	99/810	155/810	0.072
humWC06dpc	0.089	21/810	59/810	72/810	116/810	0.048
humWC06dp	0.115	21/810	69/810	92/810	177/810	0.056
humWC06p	0.113	23/810	65/810	93/810	172/810	0.056
humWC06	0.116	24/810	64/810	92/810	183/810	0.057
(humWC06nos)	0.115	25/810	62/810	92/810	186/810	0.057
“auto bi” topics:						
humWC06dpcD	0.113	36/810	78/810	97/810	141/810	0.069
humWC06dpc	0.078	24/810	51/810	65/810	107/810	0.046
humWC06dp	0.091	24/810	55/810	73/810	148/810	0.050
humWC06p	0.088	23/810	49/810	70/810	148/810	0.048
humWC06	0.091	23/810	56/810	70/810	148/810	0.049
(humWC06nos)	0.093	22/810	56/810	72/810	153/810	0.048

Table 3: Impact of Web Techniques on Generalized Success@10 (GS10)

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
new-s	−0.007	(−0.016, 0.002)	3-10-111	−0.43 (884), −0.23 (1072), 0.10 (389)
new-p	0.018	(−0.015, 0.051)	26-24-74	0.73 (1785), 0.63 (1420), −0.54 (630)
new-d	−0.001	(−0.019, 0.018)	11-20-93	0.86 (559), 0.38 (425), −0.37 (831)
new-c	−0.009	(−0.035, 0.017)	20-4-100	−0.86 (1310), −0.79 (1163), 0.26 (831)
new-D	0.028	(0.012, 0.044)	20-0-104	0.46 (831), 0.43 (195), 0.00 (1939)
old-s	0.004	(−0.016, 0.024)	25-12-158	0.87 (227), −0.74 (217), −0.86 (1259)
old-p	0.043	(0.017, 0.070)	61-28-106	1.00 (891), 0.86 (729), −0.56 (922)
old-d	0.028	(0.010, 0.047)	36-21-138	0.80 (1467), 0.60 (398), −0.42 (1615)
old-c	0.010	(−0.011, 0.031)	43-7-145	−0.86 (865), −0.74 (516), 0.57 (794)
old-D	0.012	(0.004, 0.019)	24-0-171	0.46 (1311), 0.33 (1348), 0.00 (987)
uni-s	0.000	(−0.004, 0.005)	12-12-786	−0.79 (1336), 0.55 (52), 0.74 (374)
uni-p	−0.003	(−0.007, 0.002)	36-80-694	0.63 (1023), −0.41 (1410), −0.63 (1833)
uni-d	0.002	(−0.002, 0.006)	48-50-712	0.72 (102), 0.59 (440), −0.42 (374)
uni-c	−0.025	(−0.036,−0.015)	42-66-702	−0.93 (1417), −0.93 (235), 0.54 (1833)
uni-D	0.034	(0.023, 0.045)	78-0-732	1.00 (1408), 1.00 (1252), 0.00 (1937)
bi-s	−0.002	(−0.007, 0.003)	14-12-784	0.94 (569), −0.68 (265), −0.73 (898)
bi-p	−0.002	(−0.007, 0.002)	24-58-728	1.00 (154), −0.63 (569), −0.64 (1580)
bi-d	0.003	(−0.001, 0.006)	43-39-728	0.77 (141), 0.27 (1690), −0.38 (167)
bi-c	−0.013	(−0.021,−0.005)	45-50-715	−0.93 (548), −0.93 (1005), 0.70 (1076)
bi-D	0.035	(0.024, 0.046)	68-0-742	1.00 (383), 1.00 (1390), 0.00 (1940)

- The ‘d’ technique (modest extra weight for less deep urls), which in the past has been beneficial for home page queries though at best neutral for named page queries, was neutral on average for the new and automatically-generated queries this year. This year’s new manually-created queries were said to be mainly named page; we have not yet checked if the selection of last year’s manually-created queries includes home page queries or not.

- The ‘c’ technique (duplicate filtering) was usually successful on the manually-created queries. While the few downsides on particular topics are large, in our experience, this is usually from the official judgements failing to mark all of the duplicates (though we have not checked for this year’s queries yet).

- The ‘D’ technique (domain filtering), as expected, never caused the score to go down on any topic (as the ‘vs.’ column shows) because it just included rows from the known domain. But the benefit was not large on average, so apparently the unfiltered queries usually were not confused much by the extra domains. The benefits were bigger for the automatically-generated topics, suggesting perhaps that they used less discriminative terms.

References

[1] AltaVista’s Babel Fish Translation Service. http://babelfish.altavista.com/tr

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

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

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

[8] Börkur Sigurbjörnsson, Jaap Kamps and Maarten de Rijke. EuroGOV: Engineering a Multilingual Web Corpus. *Working Notes of CLEF 2005.*

[9] Börkur Sigurbjörnsson, Jaap Kamps and Maarten de Rijke. Overview of WebCLEF 2005. *Working Notes of CLEF 2005.*

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

[11] Stephen Tomlinson. Comparing the Robustness of Expansion Techniques and Retrieval Measures. To appear in *Working Notes of CLEF 2006.*

[12] Stephen Tomlinson. European Web Retrieval Experiments with Hummingbird SearchServer$^{\text{TM}}$ at CLEF 2005. *Working Notes of CLEF 2005.*

[13] Stephen Tomlinson. Experiments in Named Page Finding and Arabic Retrieval with Hummingbird SearchServer$^{\text{TM}}$ at TREC 2002. *Proceedings of TREC 2002.*

[14] Stephen Tomlinson. Robust, Web and Genomic Retrieval with Hummingbird SearchServer$^{\text{TM}}$ at TREC 2003. *Proceedings of TREC 2003.*

[15] Stephen Tomlinson. Robust, Web and Terabyte Retrieval with Hummingbird SearchServer$^{\text{TM}}$ at TREC 2004. *Proceedings of TREC 2004.*

[16] Thijs Westerveld, Wessel Kraaij and Djoerd Hiemstra. Retrieving Web Pages using Content, Links, URLs and Anchors. *Proceedings of TREC 2001.*