European Ad Hoc Retrieval Experiments with Hummingbird SearchServer™ at CLEF 2005

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

August 21, 2005

Abstract

Hummingbird participated in the 4 monolingual information retrieval tasks (Bulgarian, French, Hungarian and Portuguese) of the Ad-Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2005. In the ad hoc retrieval tasks, the system was given 50 natural language queries, and the goal was to find all of the relevant documents (with high precision) in a particular document set. We conducted diagnostic experiments with different techniques for matching word variations and handling stopwords. We found that the experimental stemmers significantly increased mean average precision for the 4 languages. Analysis of individual topics found that the algorithmic Bulgarian and Hungarian stemmers encountered some unanticipated stopword collisions. A comparison to an experimental 4-gram technique suggested that Hungarian stemming would further benefit from decompounding. A blind feedback technique which significantly increased mean average precision for some languages was also significantly detrimental to the rank of the first relevant retrieved for one language.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance, Experimentation

Keywords

Bulgarian Retrieval, Hungarian Retrieval, First Relevant Score, Per-Topic Analysis

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],

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

Table 1: Sizes of CLEF 2005 Ad-Hoc Track Test Collections

Language	Text Size (uncompressed)	Documents	Topics	Rel/Topic
Portuguese	591,987,753 bytes	210,734	50	58 (lo 2, med 44, hi 239)
French	508,863,606 bytes	177,452	50	51 (lo 1, med 35, hi 185)
Bulgarian	216,432,023 bytes	69,195	49	16 (lo 1, med 10, hi 69)
Hungarian	106,631,823 bytes	49,530	50	19 (lo 1, med 13, hi 87)

NTCIR [6] and TREC [11]) 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 relevant documents for natural language queries in 4 European languages (Bulgarian, French, Hungarian and Portuguese) using the CLEF 2005 Ad-Hoc Track test collections.

2 Methodology

2.1 Data

The CLEF 2005 Ad-Hoc Track document sets consisted of tagged (SGML-formatted) news articles in 4 different languages: Bulgarian, French, Hungarian and Portuguese. Table 1 gives the sizes.

The CLEF organizers created 50 natural language "topics" (numbered 251-300) and translated them into many languages. One topic was discarded for Bulgarian because it had no relevant documents. Table 1 gives the final number of topics for each language and their average number of relevant documents (along with the lowest, median and highest number of relevant documents of any topic). For more information on the CLEF test collections, see the track overview paper.

2.2 Indexing

Our indexing approach was the mostly the same as last year [15]. Accents were not indexed except for the combining breve in Bulgarian. The apostrophe was treated as a word separator for the 4 investigated languages. The custom text reader, cTREC, was updated to maintain support for the CLEF guidelines of only indexing specifically tagged fields.

Some stop words were excluded from indexing (e.g. "the", "by" and "of" in English). For these experiments, the stop word list for Portuguese was based on the Porter list [7], and the lists for Bulgarian and Hungarian were based on Savoy's [9]. We used our own list for French.

Unlike previous years, this year we added AL="0-9" to the stopfiles to specify that the digits 0-9 were to be treated as alphabet characters (e.g. so that "G7" would be indexed as 1 term instead of 2).

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).

For many languages (including French and Portuguese), SearchServer provides the option of finding inflections based on lexical stemming (i.e. stemming based on a dictionary or lexicon for the language). For example, in English, "baby", "babied", "babies", "baby's" and "babying" all have "baby" as a stem. Specifying an inflected search for any of these terms will match all of the others. The lexical stemming of the post-6.0 experimental development version of SearchServer used for the experiments in this paper was based on internal stemming component 3.7.0.15. We treat each linguistic component as a black box in this paper.

Lexical stemming in SearchServer typically does "inflectional" stemming which generally retains the part of speech (e.g. a plural of a noun is typically stemmed to the singular form). It typically does not do "derivational" stemming which would often change the part of speech or the meaning more substantially (e.g. "performer" is not stemmed to "perform").

Lexical stemming in SearchServer includes compound-splitting (decompounding) for compound words in particular languages (such as Dutch, Finnish, German and Swedish). For example, in German, "babykost" (baby food) has "baby" and "kost" as stems.

Lexical stemmers can produce more than one stem, even for non-compound words. For example, in English, "axes" has both "axe" and "axis" as stems (different meanings), and in French, "important" has both "important" (adjective) and "importer" (verb) as stems (different parts of speech). SearchServer records all the stem mappings at index-time to support maximum recall and does so in a way to allow searching to weight some inflections higher than others.

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 Bulgarian topic 279 whose Title was "Референдуми в Швейцария" (Swiss referendums), a corresponding SearchSQL query would be:

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM CLEF05BG
WHERE FT_TEXT CONTAINS 'Референдуми'|'в'|'Швейцария'
ORDER BY REL DESC;
```

(Note that "в" is a stopword for Bulgarian so its inclusion in the query wouldn't actually add any matches.)

Most aspects of the SearchServer relevance value calculation are the same as described last year [15]. Briefly, SearchServer dampens the term frequency and adjusts for document length in a manner similar to Okapi [8] and dampens the inverse document frequency using an approximation of the logarithm. These calculations are based on the stems of the terms (roughly speaking) when doing morphological searching (i.e. when SET TERM_GENERATOR 'word!ftelp/inflect' was previously specified). The SearchServer RELEVANCE_METHOD setting was set to '2:3' and RELEVANCE_DLEN_IMP was set to 750 for all experiments in this paper.

2.4 Diagnostic Runs

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

- "lex": (FR and PT only): The run used SearchServer lexical stemming. The /inflect option (SET TERM_GENERATOR 'word!ftelp/inflect') was specified.
- "lexnos": Same as "lex" except that /nostop was additionally specified which prevents query terms from being discarded if all of their stems are stopwords (note that stopwords themselves were still not found because they were not indexed).
- "lexall": Same as "lex" except that a separate index was used which did not stop any words from being indexed (specifying /nostop would make no difference with this index).
- "lexsing": Same as "lex" except that /single was additionally specified (so that just one stemming interpretation was used at search time).
- "neu" (BG and HU only): Same as "lex" except that the experimental Neuchatel stemmer was used [9].
- "neunos": Same as "lexnos" except that the Neuchatel stemmer was used.
- "neuall": Same as "lexall" except that the Neuchatel stemmer was used.

Table 2: Mean Scores of Diagnostic Title-only runs

Run	FRS	Success@1	Success@5	Success@10	MRR	MAP
BG-neuall	0.782	15/49 (31%)	38/49 (78%)	41/49 (84%)	0.500	0.255
BG-neunos	0.781	16/49 (33%)	38/49 (78%)	41/49 (84%)	0.507	0.263
BG-4gram	0.758	20/49 (41%)	32/49 (65%)	40/49 (82%)	0.525	0.264
BG-snru	0.757	17/49 (35%)	34/49 (69%)	40/49 (82%)	0.499	0.242
BG-neu	0.749	15/49 (31%)	35/49 (71%)	39/49 (80%)	0.476	0.259
BG-none	0.685	14/49 (29%)	30/49 (61%)	35/49 (71%)	0.440	0.195
FR-sn	0.820	27/50 (54%)	40/50 (80%)	43/50 (86%)	0.645	0.318
FR-lex	0.810	25/50 (50%)	39/50 (78%)	42/50 (84%)	0.618	0.302
FR-lexnos	0.810	25/50 (50%)	39/50 (78%)	42/50 (84%)	0.618	0.302
FR-lexall	0.810	25/50 (50%)	39/50 (78%)	43/50 (86%)	0.618	0.301
FR-4gram	0.809	24/50 (48%)	41/50 (82%)	43/50 (86%)	0.617	0.279
FR-lexsing	0.802	25/50 (50%)	39/50 (78%)	42/50 (84%)	0.615	0.299
FR-none	0.778	20/50 (40%)	38/50 (76%)	43/50 (86%)	0.549	0.232
HU-4gram	0.834	24/50 (48%)	39/50 (78%)	45/50 (90%)	0.619	0.341
HU-neunos	0.789	26/50 (52%)	36/50 (72%)	42/50 (84%)	0.625	0.287
HU-neuall	0.788	25/50 (50%)	37/50 (74%)	41/50 (82%)	0.614	0.280
HU-neu	0.788	25/50 (50%)	37/50 (74%)	42/50 (84%)	0.613	0.274
HU-neuposs	0.769	24/50 (48%)	36/50 (72%)	41/50 (82%)	0.588	0.271
HU-none	0.671	17/50 (34%)	30/50 (60%)	37/50 (74%)	0.464	0.184
PT-sn	0.892	30/50 (60%)	43/50 (86%)	47/50 (94%)	0.712	0.269
PT-lexall	0.865	30/50 (60%)	42/50 (84%)	46/50 (92%)	0.707	0.300
PT-lex	0.856	31/50 (62%)	42/50 (84%)	45/50 (90%)	0.714	0.300
PT-lexnos	0.856	31/50 (62%)	42/50 (84%)	45/50 (90%)	0.714	0.300
PT-lexsing	0.843	30/50 (60%)	40/50 (80%)	44/50 (88%)	0.699	0.290
PT-none	0.821	28/50 (56%)	39/50 (78%)	43/50 (86%)	0.662	0.246
PT-4gram	0.815	27/50 (54%)	41/50 (82%)	41/50 (82%)	0.662	0.231

- "neuposs" (HU only): Same as "neu" except that the call to the remove_possessive function was skipped. (Prof. Savoy suggested to us that it was unclear if removing possessive pronouns was a good idea, which we interpreted as uncertainty about the remove_possessive function.)
- "sn" (FR and PT only): Same as "lex" except that the Porter (Snowball) stemmer [7] was used.
- "snru" (BG only): Same as "neu" except that the Porter (Snowball) stemmer for Russian was used.
- "4gram": Same as "lexall" except that the run used a different index which primarily consisted of the 4-grams of terms, e.g. the word 'search' would produce index terms of 'sear', 'earc' and 'arch'. No stemming was done; searching used the IS_ABOUT predicate (instead of the CONTAINS predicate) with morphological options disabled to search for the 4-grams of the query terms.
- "none": The run disabled morphological searching. (The run used the same index as "lex" for FR and PT and the same index as "neu" for HU and BG, but SET TERM_GENERATOR '' was specified so that variations from stemming were not matched.)

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

2.5 Evaluation Measures

Traditionally in ad hoc retrieval experiments, the primary evaluation measure is "average precision". For a topic, it 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, it is based on the first 1000 retrieved documents for the topic. The score ranges from 0.0 (no relevants found) to 1.0 (all relevants found at the top of the list). Average precision takes into account both precision and recall, and it is very good for detecting retrieval differences because even small differences in the ranks of relevant documents affect the score. "Mean Average Precision" (MAP) is the mean of the average precision scores over all of the topics (i.e. all topics are weighted equally).

If one wishes to focus on just the first relevant document, the traditional measure is "Reciprocal Rank" (RR). For a topic, it 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.

An experimental measure introduced in this paper (along with the companion web retrieval paper [12]) is "First Relevant Score" (denoted "FRS"). Like reciprocal rank, it is based on just the rank of the first relevant retrieved for a topic, but it is better suited to per-topic analysis. FRS 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. Like reciprocal rank, finding the first relevant at rank 1 produces a score of 1.0. At rank 2, FRS is just 7 points lower (0.93), whereas RR is 50 points lower (0.50). At rank 3, FRS is another 7 points lower (0.86), whereas RR is 17 points lower (0.33). At rank 10, FRS is 0.50, whereas RR is 0.10. FRS is greater than RR for ranks 2 to 52 and lower for ranks 53 and beyond. A possible interpretation of FRS is that it may be an indicator of the percentage of potential result list reading the system saved the user to get to the first relevant, assuming that users are less and less likely to continue reading as they get deeper into the result list.

"Success@n" is the percentage of topics for which at least one relevant document was returned in the first n rows. Like the other first relevant measures, this measure hides a lot of retrieval differences (particularly in recall), but it is more intuitive and may be an indicator of a user's impression of a method's robustness across topics. This paper lists Success@1, Success@5 and Success@10.

2.6 Statistical Significance Tables

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

- "Expt" specifies the experiment. The language code is given, followed by the labels of the 2 runs being compared. The difference is the first run minus the second run. For example, "FR lex-none" specifies the difference of subtracting the scores of the French 'none' run from the French 'lex' run (of Table 2).
- "ΔMAP" is the difference of the mean average precision scores of the two runs being compared (and "ΔFRS" is the difference of the (mean) FRS 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 (49 for Bulgarian, 50 for the others).
- "3 Extreme Diffs (Topic)" lists 3 of the individual topic differences, each followed by the topic number in brackets (the topic numbers range from 251 to 300). 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

In the per-topic analysis, the official topic translations were used as much as possible. Online translation services were consulted at times ([5] was sometimes helpful for Hungarian, and we found the Russian-to-English translations at [1] often worked for Bulgarian). Prof. Savoy also assisted with some Bulgarian words. But any translation errors are the responsibility of the author.

3.1 Impact of Stemming

Table 3 isolates the impact of stemming on the average precision measure (e.g. "FR lex-none" is the difference of the "FR-lex" and "FR-none" runs of Table 2). For each of the 4 languages, the increase in mean average precision was statistically significant (i.e. zero was not in the approximate 95% confidence interval). In FRS, there was higher variance, and only the increase for Hungarian was statistically significant. Note that for some queries, it was still better to only match the original query form (not variations from stemming); SearchServer allows this option to be controlled for each query term at search-time.

Table 3 shows that topic 279 (Swiss referendums) was substantially affected by stemming for all 4 languages, so we examine it for each language:

- HU-279 (Svájci népszavazások): Without Hungarian stemming, no document contained both of the query terms. No relevant document contained the query word 'népszavazások'. Only some of the relevant documents even contained 'Svájci' (and lots of non-relevants also did). With stemming, average precision was 87 points higher from extra matches such as 'svájciak', 'Svájc', 'Svájcban', 'Svájcot', 'Svájcról', 'népszavazáson', 'népszavazás', 'népszavazást' and 'népszavazással'.
- BG-279 (Референдуми в Швейцария): With Bulgarian stemming, average precision was 58 points higher from extra matches for 'referendums' such as референдум and референдума.

Table 3: Impact of Stemming on Average Precision and First Relevant Score

Expt	ΔMAP	95% Conf	vs.	3 Extreme Diffs (Topic)
HU-neu-none	0.090	(0.038, 0.143)	32-11-7	0.87 (279), 0.77 (294), −0.12 (265)
FR-lex-none	0.070	(0.028, 0.112)	29-16-5	0.53 (297), 0.45 (284), −0.12 (275)
BG-neu-none	0.064	(0.005, 0.123)	29-15-5	0.90 (271), 0.58 (279), −0.50 (258)
PT-lex-none	0.054	(0.027, 0.080)	34-13-3	0.35 (279), 0.30 (286), −0.09 (296)
	ΔFRS			
HU-neu-none	0.117	(0.024, 0.209)	19-10-21	1.00 (271), 0.98 (294), −0.83 (262)
BG-neu-none	0.064	(−0.042, 0.170)	16-17-16	0.96 (294), 0.86 (269), −0.87 (273)
PT-lex-none	0.035	(−0.017, 0.087)	12-7-31	0.69 (263), 0.60 (254), −0.54 (282)
FR-lex-none	0.033	(−0.032, 0.097)	15-8-27	0.73 (276), 0.64 (284), −0.60 (279)

Table 4: Impact of /nostop Option on Average Precision and First Relevant Score

Expt	ΔMAP	95% Conf	vs.	3 Extreme Diffs (Topic)
HU-nos-neu	0.013	(−0.005, 0.031)	3-1-46	0.40 (292), 0.13 (265), −0.03 (282)
BG-nos-neu	0.005	(−0.003, 0.012)	2-2-45	0.17 (273), 0.06 (267), −0.01 (257)
FR-nos-lex	0.000	n/a	0-0-50	0.00 (276), 0.00 (252), 0.00 (300)
PT-nos-lex	0.000	n/a	0-0-50	0.00 (276), 0.00 (252), 0.00 (300)
	ΔFRS			
BG-nos-neu	0.031	(−0.010, 0.072)	3-1-45	0.80 (273), 0.57 (267), −0.05 (257)
HU-nos-neu	0.001	(−0.014, 0.015)	1-1-48	0.26 (292), 0.00 (253), −0.23 (282)

- PT-279 (Referendos suíços): The query word 'suíços' was common in the relevant documents, but many relevants just used 'referendo' and not the query word 'referendos'. Average precision was 35 points higher with Portuguese stemming; extra matches included 'referendo', 'suíço', 'suíça' and 'suíças'.
- FR-279 (Référendums en Suisse): This French topic scored lower with stemming (the rank of the first relevant fell from 1 to 13, and average precision fell from 0.10 to 0.01). It appears that the relevant documents were more likely to use the plural 'Référendums' than the singular 'Référendum', and the latter was a more common word which generated lots of matches when stemming.

3.2 Impact of Experimental /nostop Option

Table 4 isolates the impact of using the SearchServer /nostop option. The option had no effect on the 50 French and Portuguese topics, and it affected only a few of the Bulgarian and Hungarian topics. The /nostop option prevents query terms from being discarded if all of their stems are stopwords (note that stopwords themselves are still not found because they are not indexed). The default is to not use /nostop because past experiments otherwise found a lot of spurrious matches in some languages (such as Finnish and Korean). We investigate some of the topics flagged in Table 4:

- HU-265 (A Deutsche Bank szerzeményei (Deutsche Bank Takeovers)): The query word 'Bank' stemmed to 'ban' (in) which was a stopword, so by default, the word 'Bank' was not matched in the documents. With the /nostop option, 'Bank' was matched and average precision was 13 points higher. (Incidentally, this issue is presumably why Table 3 shows that stemming scored 12 points lower on HU-265; without stemming, 'Bank' was found in

the documents.) Perhaps this issue would not have arisen with a lexical stemmer which would preserve the meaning more closely.

- HU-292 (Német városok újjáépítése (Rebuilding German Cities)): The query word 'Német' (German) stemmed to 'nem' (not) which was a stopword and so this useful word was dropped from the query by default. With the /nostop option, average precision was 40 points higher.
- HU-282 (Elítéltekkel szembeni durva bánásmód (Prison Abuse)): In this topic, the default scored higher. Using /nostop changed the rank of the first relevant from 3 to 7. The stopword list contained 'szemben' (in front of), and the query word 'szembeni' presumably is a related noise word, and discarding it was useful. The /nostop option kept 'szembeni', which only occurred in 319 documents, so it had a high enough weighting from inverse document frequency to hurt precision.
- BG-273 (Разширяването на НАТО (NATO Expansion)): НАТО (NATO) stemmed to НА (on) which was a stopword, so the default behaviour removed a key word from the query. With /nostop, the first relevant score was 80 points higher.
- BG-267 (Най-добрите чуждоезикови филми (Best Foreign Language Films)): The query word филми (films) stemmed to филм (film) which surprisingly was a stopword, so the default behaviour discarded a key query term. Our supplier [9] has confirmed that this was an error in the Bulgarian stopword list.
- BG-257 (Етническото прочистване на Балканите (Ethnic Cleansing in the Balkans)): The query word Балканите (Balkans) stemmed to балкан (Balkan mountain) which surprisingly was a stopword. Even though it turned out that precision was a little higher without the Balkans term in this case, in general this appears to be another error in the stopword list.

In the topics we examined, in 3 cases the default behaviour of dropping useful terms may have been from the stemmers for Bulgarian and Hungarian being algorithmic instead of lexical (a lexical stemmer typically does not change the meaning of a word, except when words are ambiguous). It appears for algorithmic stemmers it may be better to use the /nostop option by default.

In another 2 cases, it appears the stoplist was in error, which illustrates the usefulness of the CLEF judged test collections: they enable an analyst who does not understand a language to find issues in a resource for the language and make inferences about its quality.

3.3 Impact of Indexing All Words

Table 5 isolates the impact of indexing all words (i.e. of not using a stopword list). None of the mean differences were statistically significant, but Bulgarian and Hungarian had some large per-topic differences in average precision which we investigate:

- HU-292 (Német városok újjáépítése (Rebuilding German Cities)): We saw earlier that this topic benefitted from the /nostop option (average precision up 40 points), but when indexing all words, average precision fell back (33 points). The reason was that the common word 'nem' (not) was now indexed, so 'Német' (German), which stems to 'nem' with the algorithmic stemmer, had a much lower inverse document frequency than before, and this useful word received less weight. (Even if it had received more weight, there would have been potential confusion with all the indexed occurrences of 'nem'.)
- BG-271 (Бракове между хомосексуални (Gay Marriages)): The stopword между (between) was not in the 2 relevant documents. When it was indexed, its inclusion caused some non-relevants to be preferred, and average precision dropped 55 points.
- BG-295 (Пране на пари (Money Laundering)): This topic scored higher when indexing all words. Surprisingly, the word пари (money) was a stopword, presumably another error (the Bulgarian stoplist apparently needs a review). It seems fine that на (on) was a stopword.

Table 5: Impact of Indexing All Words on Average Precision and First Relevant Score

Expt	ΔMAP	95% Conf	vs.	3 Extreme Diffs (Topic)
PT-all-nos	−0.000	(−0.003, 0.002)	18-17-15	0.03 (280), −0.01 (259), −0.02 (282)
FR-all-nos	−0.001	(−0.005, 0.003)	24-17-9	−0.07 (262), 0.01 (290), 0.01 (289)
HU-all-nos	−0.006	(−0.021, 0.008)	7-7-36	−0.33 (292), −0.05 (265), 0.05 (274)
BG-all-nos	−0.008	(−0.034, 0.018)	16-17-16	−0.55 (271), −0.14 (268), 0.20 (295)
	ΔFRS			
PT-all-nos	0.009	(−0.007, 0.025)	5-1-44	0.38 (282), 0.06 (263), −0.07 (291)
BG-all-nos	0.001	(−0.008, 0.010)	3-4-42	0.13 (263), −0.07 (268), −0.07 (271)
FR-all-nos	−0.000	(−0.009, 0.008)	4-4-42	0.10 (286), −0.09 (258), −0.09 (288)
HU-all-nos	−0.000	(−0.010, 0.009)	1-3-46	0.16 (282), −0.04 (299), −0.14 (292)

Table 6: 4-grams vs. Stems in Average Precision and First Relevant Score

Expt	ΔMAP	95% Conf	vs.	3 Extreme Diffs (Topic)
HU-4gr-all	0.060	(0.018, 0.103)	32-17-1	0.46 (255), 0.33 (292), −0.30 (283)
BG-4gr-all	0.009	(−0.028, 0.046)	25-24-0	0.50 (258), 0.25 (254), −0.33 (285)
FR-4gr-all	−0.021	(−0.048, 0.005)	18-31-1	0.25 (291), 0.22 (263), −0.20 (273)
PT-4gr-all	−0.068	(−0.104,−0.032)	14-35-1	−0.43 (259), −0.28 (286), 0.22 (297)
	ΔFRS			
HU-4gr-all	0.046	(−0.036, 0.128)	15-15-20	1.00 (286), 0.93 (261), −0.81 (251)
FR-4gr-all	−0.001	(−0.041, 0.039)	13-15-22	0.60 (279), 0.26 (281), −0.40 (259)
BG-4gr-all	−0.024	(−0.093, 0.045)	17-14-18	−0.82 (274), 0.56 (270), 0.59 (288)
PT-4gr-all	−0.051	(−0.134, 0.032)	7-17-26	−1.00 (259), −0.83 (292), 0.96 (260)

In practice, indexing all words may not be so troublesome because it is typically easy for users to omit noise words from the query, and stemming issues can be worked around by disabling the finding of word variants (SearchServer makes it optional at search-time).

3.4 Comparison to 4-grams

Compound words appear to be fairly common in Hungarian, but the algorithimic stemmer did not perform decompounding, a technique we have found to be useful for languages such as Finnish [15]. However, [4] has found that using 4-grams as index terms works well in ad hoc ranking experiments for many European languages, including compound-word languages. Table 6 compares our 4-gram runs to the stemming runs which indexed all words (because we did not use stopwords with our 4-gram index). As anticipated, there was a statistically significant increase in mean average precision for Hungarian, though there was a decrease for Portuguese which was also statistically significant. We look at the largest per-topic differences for Hungarian:

- HU-255 (Internetfüggők (Internet Junkies)): Average precision was 46 points higher with 4-grams for this topic (a compound word). The stemmer found the 3 relevant documents which contained 'internetfüggő' or the original query word 'internetfüggők'. 4-grams matched other variants such as 'Internetfüggőség' (Internet dependence), 'internetfüggőséggel' and 'internetfüggőségben' and found all 6 relevant documents. 4-grams also matched other potentially helpful words such as 'internet', 'internetezők', 'internetezés', 'komputerfüggőséget' and 'függővé'. But 4-grams also produced unwanted matches, such as 'intervallum' (interval) and 'Szinte' (as good as); these both came from the 4-gram 'inte'. If the stemmer had just additionally matched 'Internetfüggőség', all 6 relevants would have found, but we're still

investigating if the -seg suffix is one that a Hungarian stemmer should generally remove or not.

- HU-292 (Német városok újjáépítése (Rebuilding German Cities)): On this topic, 4-grams still just found 1 of the 2 relevant documents, but it moved it from rank 3 to 1 (compared to the stemming run). While 4-grams additionally matched 'újjáépítik', the bigger advantage was probably that the 4-gram method did not match 'nem' which we know from earlier was a troublesome match for the stemming run.

- HU-283 (James Bond-filmek (James Bond Films)): On this topic, the 4-gram run scored 30 points lower in average precision than the stemming run. The 4-gram run favored documents with the 'filmek' pattern (which corresponded to three 4-grams ('film', 'ilme' and 'lmek') and so it received roughly 3 times the weight compared to the stemming run). However, the relevant documents tended not to use 'filmek'; instead they tended to use other variants matched by the stemmer such as 'film', 'filmet', 'filmnél', 'filmben' and 'filmhez'.

- HU-286 (Futballsérülések (Football Injuries)): This topic had no matches in the stemming run, but a relevant document was ranked first in the 4-gram run. 4-gram matches in the relevant documents included 'futballista', 'futballkapus' (goalkeeper), 'futballválogatott', 'vállsérülést', 'vállsérüléssel', 'vállsérülés', 'sérülés' (injury), 'sérült' and 'sérültet'. This might be a case for which decompounding would be helpful.

- HU-261 (Jövendőmondás (Fortune-telling)): The stemming run only matched the one document which contained 'jövendőmondást' and 'jövendőmondás' and it was judged non-relevant, so it scored 0 on this topic. The 4-gram returned 1 of the 3 relevant documents at rank 2 (the others weren't ranked in the top 100). Matches in the relevant document included 'jövendölők' and 'jövendőmondók'. The latter of these perhaps could have been matched with additional stemming rules, but the former would require a stemmer to do decompounding (or, if the user had decompounded the query, the latter would require index-time decompounding to match).

SearchServer can find character sequences inside European words without n-gramming if the user specifies wildcards, so for precise searches it's unclear if n-gram indexes would add value. N-gram approaches typically produce larger indexes and its queries can be slower for common word-searching cases. We're not aware of them being used in practice for European language retrieval, except perhaps by web search engines for url indexing.

3.5 Comparison to Alternate Stemmers

Table 7 compares alternate stemming approaches to the approach we used in our submitted runs. Unfortunately, we have run out of time to examine more topics in detail for this draft paper, but we note in particular that it seems not to matter very much on average whether the remove_possessive function of the Hungarian stemmer is called or not.

3.6 Impact of /single Option

Table 8 isolates the impact of using the SearchServer /single option. This option only makes a difference for the SearchServer lexical stemmers which can produce more than one stem for a term. Like last year [15], our method for including all stems without overweighting some of the terms apparently was effective. Even in the high-variance first relevant score measure, the bigger differences favored including all stems.

Table 7: Alternate Stemming vs. Baseline in Average Precision and First Relevant Score

Expt	ΔMAP	95% Conf	vs.	3 Extreme Diffs (Topic)
FR-sn-lex	0.017	(0.001, 0.032)	20-16-14	0.29 (291), 0.15 (287), −0.08 (278)
HU-poss-neu	−0.003	(−0.017, 0.012)	18-9-23	−0.27 (268), 0.11 (258), 0.13 (262)
BG-snru-neu	−0.017	(−0.064, 0.029)	19-25-5	−0.64 (259), −0.44 (271), 0.50 (258)
PT-sn-lex	−0.031	(−0.060,−0.001)	21-23-6	−0.41 (279), −0.28 (286), 0.21 (274)
	ΔFRS			
PT-sn-lex	0.036	(−0.024, 0.096)	10-8-32	0.96 (260), 0.49 (300), −0.59 (292)
FR-sn-lex	0.010	(−0.005, 0.025)	7-7-36	0.19 (252), 0.16 (299), −0.12 (251)
BG-snru-neu	0.008	(−0.070, 0.086)	14-13-22	0.87 (273), 0.84 (270), −0.79 (280)
HU-poss-neu	−0.019	(−0.078, 0.040)	4-5-41	−0.95 (265), −0.68 (270), 0.69 (262)

Table 8: Impact of /single Option on Average Precision and First Relevant Score

Expt	ΔMAP	95% Conf	vs.	3 Extreme Diffs (Topic)
FR-sing-lex	−0.002	(−0.011, 0.007)	8-7-35	−0.15 (297), −0.10 (284), 0.11 (263)
PT-sing-lex	−0.010	(−0.018,−0.002)	8-11-31	−0.10 (292), −0.10 (275), 0.02 (298)
	ΔFRS			
FR-sing-lex	−0.009	(−0.025, 0.008)	1-2-47	−0.40 (259), −0.06 (284), 0.03 (299)
PT-sing-lex	−0.013	(−0.037, 0.011)	1-3-46	−0.59 (292), −0.07 (275), 0.06 (267)

Table 9: Mean Scores of Submitted Runs

Run	FRS	Success@1	Success@5	Success@10	MRR	MAP
humBG05t	0.749	15/49 (31%)	35/49 (71%)	39/49 (80%)	0.476	0.259
humBG05td	0.815	18/49 (37%)	39/49 (80%)	42/49 (86%)	0.537	0.275
humBG05tde	0.752	21/49 (43%)	35/49 (71%)	38/49 (78%)	0.549	0.298
humFR05t	0.810	25/50 (50%)	39/50 (78%)	42/50 (84%)	0.618	0.302
humFR05td	0.825	30/50 (60%)	39/50 (78%)	41/50 (82%)	0.686	0.369
humFR05tde	0.822	31/50 (62%)	40/50 (80%)	41/50 (82%)	0.697	0.401
humHU05t	0.788	25/50 (50%)	37/50 (74%)	42/50 (84%)	0.613	0.274
humHU05td	0.838	23/50 (46%)	41/50 (82%)	43/50 (86%)	0.614	0.306
humHU05tde	0.835	22/50 (44%)	38/50 (76%)	45/50 (90%)	0.602	0.331
humPT05t	0.856	31/50 (62%)	42/50 (84%)	45/50 (90%)	0.714	0.300
humPT05td	0.939	35/50 (70%)	48/50 (96%)	49/50 (98%)	0.805	0.357
humPT05tde	0.925	35/50 (70%)	47/50 (94%)	48/50 (96%)	0.799	0.386

4 Submitted Runs

Table 9 lists the mean scores of the runs submitted for assessment in May 2005. In the identifiers (e.g. "humFR05tde"), '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 2 rows was used (see the 2003 paper [14] for more details). From the Description fields for Bulgarian, French and Portuguese, instruction words such as "find", "relevant" and "document" were automatically removed (based on looking at some older topic lists, not this year's topics; this step was skipped for Hungarian because we lacked an older topic list).

The submitted French and Portuguese Title-only runs (i.e. "humFR05t" and "humPT05t" of Table 9) correspond to the "lex" diagnostic runs (i.e. "FR-lex" and "PT-lex" of Table 2) except that the submitted runs used an older experimental version of SearchServer (though there don't appear to have been any differences that affected the runs). The submitted Bulgarian and Hungarian Title-only runs (i.e. "humBG05t" and "humHU05t"of Table 9) correspond to the "neu" diagnostic runs (i.e. "BG-neu" and "HU-neu" of Table 2).

4.1 Impact of Adding the Description Field

Table 10 isolates the impact of adding the Description field to the query. Though adding the Description tended to increase the scores on average (and in some cases this result was statistically significant), one should keep in mind that the Description often repeated the Title words, which hence received twice the weight in the combined query. We would expect to see more variance if the Title was replaced by the Description instead of being augmented by it

4.2 Impact of Blind Feedback

Table 11 isolates the impact of the blind feedback technique (based on using the first 2 returned rows to expand the query). While mean average precision increased for all 4 languages (and the increase was statistically significant for 3 of them), the first relevant score decreased for all 4 languages (and the decrease was statistically significant for the other 1 of them).

The blind feedback technique presumably works best if relevant documents appear in the first 2 rows, in which case first relevant score cannot be improved. If the first 2 rows do not contain relevant documents, then using those rows to expand the query may hurt the query and push down the first relevant even further.

This result may explain in part why blind feedback techniques are not known to be used

Table 10: Impact of Description on Average Precision and First Relevant Score

Expt	ΔMAP	95% Conf	vs.	3 Extreme Diffs (Topic)
FR-td-t	0.068	(0.030, 0.105)	35-14-1	0.61 (256), 0.33 (281), −0.18 (277)
PT-td-t	0.057	(0.008, 0.107)	31-18-1	−0.45 (258), 0.34 (299), 0.34 (264)
HU-td-t	0.031	(−0.002, 0.065)	33-17-0	0.33 (286), 0.31 (290), −0.23 (274)
BG-td-t	0.016	(−0.034, 0.066)	29-19-1	−0.68 (271), −0.38 (277), 0.30 (294)
	ΔFRS			
PT-td-t	0.083	(0.005, 0.160)	18-10-22	1.00 (272), 0.86 (288), −0.50 (258)
BG-td-t	0.065	(−0.010, 0.141)	23-15-11	0.80 (273), 0.70 (286), −0.53 (278)
HU-td-t	0.049	(−0.026, 0.125)	16-16-18	1.00 (286), 0.93 (261), −0.59 (282)
FR-td-t	0.014	(−0.033, 0.062)	17-10-23	0.74 (282), −0.36 (257), −0.54 (292)

Table 11: Impact of Blind Feedback on Average Precision and First Relevant Score

Expt	ΔMAP	95% Conf	vs.	3 Extreme Diffs (Topic)
FR-tde-td	0.031	(0.015, 0.047)	34-16-0	0.17 (273), 0.16 (290), −0.07 (268)
PT-tde-td	0.029	(0.005, 0.053)	34-16-0	0.30 (290), 0.20 (275), −0.24 (274)
HU-tde-td	0.025	(0.003, 0.047)	31-17-2	0.29 (254), 0.18 (290), −0.18 (279)
BG-tde-td	0.023	(−0.002, 0.048)	29-18-2	0.50 (272), 0.14 (254), −0.10 (277)
	ΔFRS			
FR-tde-td	−0.002	(−0.041, 0.036)	10-6-34	−0.58 (282), −0.34 (272), 0.42 (252)
HU-tde-td	−0.003	(−0.037, 0.032)	7-6-37	−0.39 (298), −0.37 (300), 0.38 (269)
PT-tde-td	−0.014	(−0.038, 0.010)	6-7-37	−0.50 (258), −0.16 (277), 0.07 (269)
BG-tde-td	−0.062	(−0.109,−0.016)	9-16-24	−0.63 (277), −0.50 (299), 0.13 (296)

in practice even though they have been popular with experimenters for several years in ad hoc evaluations (which typically focus on mean average precision).

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] Paul McNamee and James Mayfield. JHU/APL Experiments in Tokenization and Non-Word Translation. *Working Notes for the CLEF 2003 Workshop*, 2003.

[5] MTA SZTAKI: English-Hungarian, Hungarian-English Online Dictionary. http://dict.sztaki.hu/english-hungarian

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

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

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

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

[10] Börkur Sigurbjörnsson, Jaap Kamps and Maarten de Rijke. Overview of WebCLEF 2005. To appear in *Working Notes for the CLEF 2005 Workshop*, 2005.

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

[12] Stephen Tomlinson. European Web Retrieval Experiments with Hummingbird SearchServer™ at CLEF 2005. To appear in *Working Notes for the CLEF 2005 Workshop*, 2005.

[13] Stephen Tomlinson. Experiments in 8 European Languages with Hummingbird SearchServer™ at CLEF 2002. *Proceedings of CLEF 2002*, 2003.

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

[15] Stephen Tomlinson. Finnish, Portuguese and Russian Retrieval with Hummingbird SearchServer™ at CLEF 2004. *Working Notes for the CLEF 2004 Workshop*, 2004.