CJK Experiments with Hummingbird SearchServer™ at NTCIR-5

Stephen Tomlinson
Hummingbird
Ottawa, Ontario, Canada
stephen.tomlinson@hummingbird.com
October 15, 2005

Abstract

Hummingbird submitted ranked result sets for the Chinese, Japanese and Korean Single Language Information Retrieval subtasks of the Cross-Lingual Information Retrieval Task of the 5th NII-NACSIS Test Collection for IR Systems Workshop (NTCIR-5). For short Chinese (title) queries, a decompounded word-based approach produced higher (statistically significant) mean average precision and first relevant scores than an overlapping n-gram approach. For Korean queries, a word-based decompounding and stemming approach produced significantly higher mean average precision scores than plain word-based matching. For Japanese title queries, a blind feedback technique which produced a statistically significant increase in mean average precision also produced a statistically significant decrease in mean first relevant score. **Keywords:** *Chinese (Traditional), Japanese, Korean, decompounding, segmenting, stemming, n-grams, 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 [4] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (NTCIR [7], CLEF [2] and TREC [9]) have provided judged test collections for objective experimentation with SearchServer in more than a dozen languages.

This paper describes experimental work with SearchServer for the task of finding relevant documents for natural language queries in 3 East Asian languages (Chinese, Japanese and Korean) using the NTCIR-5 test collections.

[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

2.1 Data

The document sets of the NTCIR-5 test collections (CLIR task) consisted of news articles from 2000 and 2001 in Chinese (Traditional), Japanese and Korean. Table 1 gives their sizes. For more details, see the CLIR task overview paper.

Table 1. Sizes of NTCIR-5 Document Sets

Language	Text Size	#Documents
Chinese	1,113,487,231 bytes	901,446
Japanese	1,078,183,238 bytes	858,400
Korean	333,320,195 bytes	220,374

The NTCIR organizers created 50 natural language "topics" (numbered 1-50) and produced a set of relevance assessments: a list of documents judged to be highly relevant, relevant, partially relevant or not relevant for each of the topics. In this paper, we just count 'highly relevant' or 'relevant' as relevant. Table 2 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 the topics).

Table 2. Judged Topics of NTCIR-5

Language	Topics	Rel/Topic
Chinese	50	38 (lo 3, med 26, hi 187)
Japanese	47	45 (lo 5, med 24, hi 293)
Korean	50	37 (lo 4, med 25, hi 153)

Table 3. Mean Scores of Diagnostic Title-only Runs

Run	FRS	S@1	S@10	MAP
C-Base-T	0.871	31/50	45/50	0.324
C-Cmpd-T	0.845	29/50	45/50	0.310
C-Ngram-T	0.807	27/50	42/50	0.290
J-Cmpd-T	0.888	28/47	45/47	0.302
J-Ngram-T	0.886	25/47	44/47	0.285
J-Base-T	0.885	28/47	44/47	0.312
K-Ngram-T	0.921	33/50	49/50	0.376
K-Nostop-T	0.916	29/50	49/50	0.358
K-Cmpd-T	0.916	27/50	49/50	0.342
K-Single-T	0.913	30/50	49/50	0.352
K-Base-T	0.912	29/50	49/50	0.355
K-None-T	0.857	28/50	46/50	0.241

Table 4. Mean Scores of Diagnostic Description-only Runs

Run	FRS	S@1	S@10	MAP
C-Base-D	0.815	24/50	43/50	0.268
C-Keep-D	0.803	23/50	43/50	0.266
C-Cmpd-D	0.780	23/50	41/50	0.261
C-Ngram-D	0.770	19/50	42/50	0.243
J-Base-D	0.814	17/47	41/47	0.281
J-Keep-D	0.799	18/47	39/47	0.272
J-Cmpd-D	0.797	19/47	40/47	0.268
J-Ngram-D	0.787	21/47	39/47	0.256
K-Keep-D	0.914	33/50	47/50	0.352
K-Base-D	0.901	33/50	46/50	0.355
K-Nostop-D	0.901	33/50	46/50	0.354
K-Single-D	0.894	36/50	45/50	0.353
K-Cmpd-D	0.894	32/50	45/50	0.343
K-Ngram-D	0.892	33/50	46/50	0.370
K-None-D	0.779	24/50	42/50	0.174

2.2 Indexing

The experimental post-6.0 version of SearchServer used in these experiments provided word-based and n-gram approaches to indexing.

Word-based approaches: For Chinese and Japanese, SearchServer segmented the text into words and optionally split compound words (decompounding). The segmenter also performed stemming for Japanese. For Korean, SearchServer indexed both the surface forms of Korean words and the stems (after decompounding). A short stopword list was used for each language. The lexicon-based segmenters and stemmers were based on internal linguistic component 3.7.0.15.

The overlapping n-gram approach (available for all 3 languages) typically used bigrams for most Asian text.

2.3 Searching

For all runs, SearchServer Intuitive Searching was used, i.e. the IS_ABOUT predicate of SearchSQL, which accepts unstructured text. For example, if the Title for a topic was "地震, 台湾" (Earthquakes, Taiwan), then a corresponding SearchSQL query would be:

```
SELECT RELEVANCE() AS REL, DOCNO
FROM NTC4J
WHERE FT_TEXT IS_ABOUT '地震, 台湾'
ORDER BY REL DESC;
```

The relevance value calculation was the mostly same as described last time [10]. Briefly, SearchServer dampened the term frequency and adjusted for document length in a manner similar to Okapi [8] and dampened the inverse document frequency using an approximation of the logarithm. This year, for the short Title-only queries, RELEVANCE_METHOD '2:3' and RELEVANCE_DLEN_IMP 250 was used. For the longer query forms (noisier queries), RELEVANCE_METHOD '2:4' (which squares the importance of inverse document frequency) and RELEVANCE_DLEN_IMP 500 was used. These settings were chosen based on experiments on older test collections. For Korean, the relevance ranking included the technique for handling multiple stemming interpretations (described in [11]).

When searching Korean with a word-based index, as of SearchServer 6.0, the user can decide at search-time for each query word whether to match if any stem matches (/inflect/decompound option), or whether to require all of its stems (from a particular stemming interpretation) to be in the same or consecutive words (/inflect option), or whether to just match on the surface form (none option), among other possibilities (explored in more detail below). The experiments in this paper always used the same option for all words of a query via the VECTOR_GENERATOR setting. (The same results could have been achieved with the CONTAINS predicate specifying a boolean-OR of the query words and a corresponding TERM_GENERATOR setting.)

A blank VECTOR_GENERATOR was used for n-gram experiments and also for word-based Chinese and Japanese experiments.

2.4 Diagnostic Runs

For the diagnostic runs listed in Tables 3 and 4, the run names start with the first letter of the language, followed by a label, followed by the topic field used

('T' for the Titles (short keyword lists) or 'D' for the Descriptions (typically one-sentence)). The labels are as follows:

"Base": The base run for Chinese and Japanese used the word-based approach with decompounding enabled. For Korean, the word-based stemming approach was used, and the search-time '/inflect/decompound' matching option was used.

"Cmpd": For Chinese and Japanese, same as Base except that a different SearchServer table was used which had decompounding mode disabled. For Korean, the same index as Base was used, but the search-time matching option was just '/inflect'.

"Ngram": Same as Base except that a different SearchServer table was used which was indexed with overlapping n-grams (and hence no special search-time matching options were available).

"Single" (Korean-only): Same as Base except '/single' was additionally specified (so that just one stemming interpretation was used at search time).

"Nostop" (Korean-only): Same as Base except '/nostop' was additionally specified which prevented query terms from being discarded if all of their stems were stopwords (note that stopwords themselves were still not found because they were not indexed).

"None" (Korean-only): Same as Base except that morphological matching was disabled via a blank VECTOR_GENERATOR (so just the surface forms were matched, not the stems).

"Keep" (Description runs only): Same as Base except that instruction words such as "find", "relevant" and "document" were not discarded before searching. The word lists for Chinese, Japanese and Korean were developed from the Descriptions of the NTCIR-3 topics.

2.5 Evaluation Measures

Traditionally in ad hoc retrieval experiments, the primary evaluation measure is "average precision" (AP). 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 has been "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 [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. 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.

Motivations for FRS: The reciprocal rank measure considers a small drop from rank 1 to 2 (50 points) to be greater than a large drop from 2 to 100 (49 points), which seems improper and causes analysis of the largest per-topic differences to be less effective at finding large retrieval differences. FRS considers a drop from rank 1 to 2 (7 points) to be much less than a drop from 2 to 100 (93 points). The choice of the 1.08 constant in FRS causes FRS to be 0.5 at rank 10 which in practice makes FRS a good predictor of Success@10 (e.g. if FRS is 0.8, Success@10 will probably be close to 40/50).

"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 (S@1) and Success@10 (S@10) for all runs.

2.6 Comparision Tables

For comparison tables such as Tables 5 and 6, the columns are as follows:

- "Expt" is a label for the experiment. When comparing diagnostic runs, the name of the non-Base run is listed.
- "ΔMAP" is the difference of the mean average precision scores when subtracting the Base run from the listed run (and "ΔFRS" is the difference of the (mean) FRS scores).
- "95% Conf" is an approximate 95% confidence interval for the mean difference (calculated from

Table 5. Mean Impact of Disabling Decompounding

Expt	ΔMAP (95% Conf)	vs.
C-Cmpd-D	−0.008 (−0.04, 0.02)	23-25-2
J-Cmpd-T	−0.010 (−0.03, 0.01)	25-22-0
K-Cmpd-D	−0.012 (−0.04, 0.01)	20-21-9
K-Cmpd-T	−0.013 (−0.04, 0.02)	16-21-13
J-Cmpd-D	−0.013 (−0.04, 0.01)	19-27-1
C-Cmpd-T	−0.014 (−0.04, 0.01)	22-27-1
	ΔFRS	
K-Cmpd-T	0.004 (−0.03, 0.04)	4-4-42
J-Cmpd-T	0.002 (−0.03, 0.04)	9-5-33
K-Cmpd-D	−0.007 (−0.04, 0.02)	7-5-38
J-Cmpd-D	−0.017 (−0.06, 0.03)	12-13-22
C-Cmpd-T	−0.026 (−0.07, 0.02)	6-9-35
C-Cmpd-D	−0.034 (−0.08, 0.01)	10-15-25

Table 6. Per-Topic Impact of Disabling Decompounding

Expt	3 Extreme AP Diffs (Topic)
C-Cmpd-D	0.42 (23), −0.23 (19), −0.34 (37)
J-Cmpd-T	−0.23 (31), −0.11 (34), 0.10 (17)
K-Cmpd-D	−0.30 (43), −0.20 (7), 0.21 (45)
K-Cmpd-T	0.32 (45), −0.24 (13), −0.29 (43)
J-Cmpd-D	−0.21 (17), −0.12 (16), 0.13 (35)
C-Cmpd-T	−0.20 (19), −0.15 (46), 0.19 (47)
	3 Extreme FRS Diffs (Topic)
K-Cmpd-T	−0.50 (13), 0.32 (7), 0.39 (44)
J-Cmpd-T	−0.46 (31), 0.21 (29), 0.36 (41)
K-Cmpd-D	−0.46 (44), −0.21 (43), 0.14 (45)
J-Cmpd-D	−0.69 (29), −0.35 (31), 0.37 (41)
C-Cmpd-T	−0.93 (33), −0.32 (19), 0.19 (47)
C-Cmpd-D	−0.61 (33), −0.46 (19), 0.19 (20)

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 score was higher, lower and tied (respectively) than the score of the base run. These numbers should always add to the number of topics (50 for Chinese and Korean, 47 for Japanese).
- "3 Extreme Diffs (Topic)" lists 3 of the individual topic differences, each followed by the topic number in brackets (the topic numbers range from 1 to 50). 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 also consulted at times (particularly [1] and sometimes [13]).

3.1 Impact of Decompounding

Table 5 shows the impact of not using decompounding on the mean average precision and FRS measures. (A negative difference means the score was higher with decompounding; note that this is the opposite convention to last year [10].) Most of the mean scores were higher when decompounding, but none of the mean differences were statistically significant.

(Last year [10] we found that decompounding produced significant gains in mean average precision for Korean. However, this year's experiment is different for Korean. Last year, we disabled decompounding at index-time. This year, we still do decompounding at index-time, but "disable" it at search-time by requiring all of the stems of a query word (from a particular stemming interpretation) to be in the same or consecutive words (i.e. internally use the phrase operator for compound words). Hence, unlike last year, a query word will match a component of an indexed compound word even without the /decompound option set.)

Decompounding had a substantial impact on some particular queries. We examine a few of them:

Topic C-019-T (超音速飛機，協和號，墜機 (supersonic airliner, Concord, airplane crash)): Table 6 shows that topic 19 had the largest decrease in average precison when disabling decompounding for Chinese title queries. Decompounding split two words of this query: 超音速 (supersonic speed) was split to 超 (ultra) and 音速 (speed of sound), and 協和號 (Concord) was split to 協和 (harmony) and 號 (number). The relevant documents apparently used the short form 協和 for Concord which only matched when decompounding.

Topic J-031-T (微細な塵埃粒子，心臓病 (fine dust particles, heart disease)): Table 6 shows that topic 31 had the largest decrease in average precison when disabling decompounding for Japanese title queries. Decompounding split the query word 心臓病 (heart disease) to 心臓 (heart) and 病 (illness). Some rele-

Table 7. Mean Impact of N-grams

Expt	ΔMAP (95% Conf)	vs.
K-Ngram-T	0.021 (−0.02, 0.06)	28-21-1
K-Ngram-D	0.014 (−0.02, 0.05)	31-18-1
C-Ngram-D	−0.025 (−0.07, 0.02)	20-30-0
J-Ngram-D	−0.025 (−0.06, 0.01)	19-28-0
J-Ngram-T	−0.027 (−0.06, 0.01)	20-27-0
C-Ngram-T	−0.034 (−0.07, 0.00)	16-33-1
	ΔFRS	
K-Ngram-T	0.010 (−0.02, 0.04)	11-6-33
J-Ngram-T	0.000 (−0.05, 0.05)	9-10-28
K-Ngram-D	−0.008 (−0.06, 0.04)	11-11-28
J-Ngram-D	−0.027 (−0.09, 0.03)	17-19-11
C-Ngram-D	−0.044 (−0.11, 0.02)	12-21-17
C-Ngram-T	−0.064 (−0.13, 0.00)	4-14-32

Table 8. Per-Topic Impact of N-grams

Expt	3 Extreme AP Diffs (Topic)
K-Ngram-T	0.34 (41), 0.33 (45), −0.19 (37)
K-Ngram-D	−0.23 (9), 0.20 (13), 0.22 (12)
C-Ngram-D	−0.43 (37), −0.31 (18), 0.34 (23)
J-Ngram-D	−0.34 (45), −0.23 (4), 0.21 (44)
J-Ngram-T	−0.41 (4), −0.40 (45), 0.16 (8)
C-Ngram-T	−0.42 (35), −0.24 (47), 0.13 (44)
	3 Extreme FRS Diffs (Topic)
K-Ngram-T	−0.41 (7), 0.26 (45), 0.32 (44)
J-Ngram-T	−0.63 (4), −0.50 (9), 0.34 (44)
K-Ngram-D	−0.73 (32), −0.37 (33), 0.39 (4)
J-Ngram-D	−0.51 (31), 0.33 (32), 0.37 (44)
C-Ngram-D	−0.59 (33), −0.54 (39), 0.50 (26)
C-Ngram-T	−0.84 (33), −0.74 (47), 0.33 (38)

vant documents did not contain the compound 心臟病 but variations such as 心臟発作 (heart seizure) which required decompounding to match.

Topic K-045-T (인구문제, 기아 (population issue, hunger)): Table 6 shows that topic 45 had the largest increase in average precison when disabling decompounding for Korean title queries. For this query, it was better for precision to keep 인구문제 (population problem) as a phrase. With the '/decompound' option, there were additional matches such as 비만인구 (obese population) which tended not to be helpful for this query. (As an aside, with either approach, SearchServer matched the Hangul word 기아 (hunger) with the Han form 飢餓 though the Han form did not occur in a relevant document.)

Topic K-043-T (2단계 핵무기감축협정(START2), 러시아, 비준 (START2, Russia, ratification)): Table 6 shows that topic 43 was the biggest beneficiary of decompounding for Korean title queries in the average precision measure. From the compound word 핵무기감축협정 (nuclear weapon reduction agreement), the '/decompound' option found several helpful matches in relevant documents that were otherwise missed such as 전략무기감축협정 (strategic weapon reduction agreement), 핵무기 감축을 (nuclear weapon reduction), 핵무기를 (nuclear weapon), 감축과 (reduction), ABM협정의 (ABM agreement), 핵공격력 (nuclear striking power), 대량살상무기 (weapons of mass destruction). A user could have found these matches without '/decompound' by splitting the compound when entering it.

Note: several more examples of CJK decompounding were in last year's paper [10].

Conclusions on Decompounding: For the CJK languages, it is commonly argued that word-based indexing should support searching on the components of compound words to increase recall (e.g. [5] prefers "short-words" to "long-words" for Chinese, [3] prefers "short unit keywords" to "long unit keywords" for Japanese, and [6] finds that "morphemes" or "simple nouns" are more effective than "words" or "compound nouns" for Korean). In this section, we measured the impact of the experimental "decompounding" options of our implementation on the NTCIR-5 CLIR ad hoc search tasks. We didn't find any statistically significant mean differences, perhaps in part because even in "non-decompounding" mode the Chinese and Japanese segmenters usually produced short words and our Korean index still used decompounded stems. In practice, SearchServer users can use a decompounded index but still enforce phrase matching at search-time for compound words when desired.

3.2 Comparison to N-grams

Table 7 shows the impact of using n-grams instead of decompounded words on the mean average precision and FRS measures. (A negative difference means the score was higher with words; note that this is the opposite convention to last year [10].) For Chinese titles, there was a statistically significant decrease in mean average precision and FRS when using n-grams. For the other cases, the mean differences were not statistically significant.

We look at some of the largest per-topic differences of n-grams and decompounded words:

Topic C-035-T (死刑，調查資料 (capital punishment, survey data)): Table 8 shows that topic 35 had the largest decrease in average precison when using n-grams for Chinese title queries. The word-based approach searched for 死刑 (death penalty), 調查 (investigation) and 資料 (material), and inverse document frequency gave more weight to 死刑 than the other two words combined. The overlapping n-gram approach added the uncommon 查資 bigram to the query which in effect more than doubled the weight on the 'survey

data' phrase. The relevant documents tended not to use that phrase, and the word-based weighting was much more effective.

Topic C-047-T (韓國大選，2000年，大國家黨 (Korean general election, 2000, Han Nara Party)): Table 8 shows that topic 47 had the next largest decrease in average precison when using n-grams for Chinese title queries. The relevant documents typically used the phrase 韓國黨 (South Korean party) which was fully matched by the word-based approach (which included 韓國 (South Korea) and 黨 (party) in the query) but was only partially matched by the n-gram approach (the bigram 韓國 matched but the only query bigram containing 黨 was 家黨 which did not match the key phrase). The n-gram approach favored matches from bigrams in the 'Han Nara Party' phrase, and also the the bigram 國大 caused some distracting matches.

Topic C-044-T (氣候異常，災害，造成 (abnormal weather, disaster, cause)): Table 8 shows that topic 44 had the largest increase in average precison when using n-grams for Chinese title queries. The word-based approach produced the key query terms of 氣候 (climate) and 異常 (exceptional), but n-grams additionally produced the uncommon bigram 候異 which in effect emphasized the 氣候異常 phrase; this phrase occurred in most of the relevant documents so this emphasis was beneficial for this query.

Topic J-004-T (米国防長官，ウィリアム・セバスチャン・コーエン，北京 (the US Secretary of Defense, William Sebastian Cohen, Beijing)): Table 8 shows that topic 4 had the largest decrease in both average precison and first relevant score when using n-grams for Japanese title queries. The n-gram approach used n>2 for some of the Katakana terms, and the 4-character word コーエン (Cohen) received the same weight as in the word-based approach. However, for the longer Katakana terms ウィリアム (William) and セバスチャン (Sebastian) the overlapping n-grams in effect assigned 3 times the weight than the word-based approach. Hence the n-gram approach favored documents containing ウィリアム (William) and セバスチャン (Sebastian) even if they were missing most of the other query terms. The relevant documents tended to refer to コーエン長官 (Secretary Cohen) or コーエン米国防長官 (Cohen, American National Defense Director) without mentioning his given names. The word-based approach returned a relevant in the first row while the n-gram approach did not return a relevant until the 14th row.

Note: several more comparisons of CJK n-grams vs. decompounded words were in last year's paper [10].

Conclusions on N-grams: Researchers have found that n-gram methods generally score comparably to the highest-scoring word-based approaches in CJK ad hoc search experiments [5, 3, 6] (though n-grams produce a larger index and have more search-time overhead). This year, we found that the word-based approach produced a (borderline) statistically significant gain in both mean average precision and first relevant score for Chinese title queries. The differences for Japanese and Korean were not statistically significant. N-gram approaches can have weighting issues when the words are of length greater than 'n'. Also, n-gram methods may not be suitable for some retrieval features (not investigated in these experiments) such as spelling-correction or supporting domain-specific thesauri.

Table 9. Mean Impact of Other Word-based Matching Options (Korean)

Expt	ΔMAP (95% Conf)	vs.
K-Nostop-T	0.003 (−0.01, 0.01)	3-1-46
K-Nostop-D	−0.001 (−0.01, 0.01)	17-20-13
K-Single-T	−0.003 (−0.03, 0.02)	17-10-23
K-Single-D	−0.003 (−0.02, 0.02)	23-18-9
K-None-T	−0.115 (−0.16,−0.07)	7-43-0
K-None-D	−0.181 (−0.24,−0.12)	7-43-0
	ΔFRS	
K-Nostop-T	0.004 (−0.01, 0.02)	1-0-49
K-Single-T	0.001 (−0.02, 0.02)	7-3-40
K-Nostop-D	0.000 n/a	0-0-50
K-Single-D	−0.007 (−0.04, 0.03)	6-4-40
K-None-T	−0.055 (−0.12, 0.01)	12-14-24
K-None-D	−0.122 (−0.23,−0.02)	9-21-20

Table 10. Per-Topic Impact of Other Word-based Matching Options (Korean)

Expt	3 Extreme AP Diffs (Topic)
K-Nostop-T	0.10 (45), 0.03 (21), −0.00 (49)
K-Nostop-D	−0.02 (43), −0.01 (9), 0.01 (45)
K-Single-T	−0.29 (35), −0.24 (40), 0.23 (21)
K-Single-D	−0.25 (40), −0.13 (45), 0.07 (9)
K-None-T	−0.47 (19), −0.43 (43), 0.13 (46)
K-None-D	−0.95 (41), −0.63 (19), 0.18 (15)
	3 Extreme FRS Diffs (Topic)
K-Nostop-T	0.19 (45), 0.00 (3), 0.00 (50)
K-Single-T	−0.21 (46), −0.16 (5), 0.14 (44)
K-Nostop-D	0.00 (26), 0.00 (2), 0.00 (50)
K-Single-D	−0.63 (45), −0.18 (40), 0.14 (44)
K-None-T	−1.00 (13), −0.79 (26), 0.33 (2)
K-None-D	−1.00 (10), −1.00 (20), 0.63 (47)

3.3 Other Word-based Matching Options (Korean)

Table 9 shows the impact of some of the other word-based matching options for Korean on the mean aver-

Table 11. Mean Impact of Keeping Instruction Words

Expt	ΔMAP (95% Conf)	vs.
C-Keep-D	−0.002 (−0.01, 0.01)	15-32-3
K-Keep-D	−0.004 (−0.02, 0.01)	17-31-2
J-Keep-D	−0.009 (−0.02, 0.00)	15-30-2
	ΔFRS	
K-Keep-D	0.013 (−0.01, 0.03)	5-2-43
C-Keep-D	−0.012 (−0.03, 0.00)	0-11-39
J-Keep-D	−0.015 (−0.04, 0.01)	6-11-30

Table 12. Per-Topic Impact of Keeping Instruction Words

Expt	3 Extreme AP Diffs (Topic)
C-Keep-D	−0.04 (14), −0.03 (16), 0.03 (46)
K-Keep-D	−0.08 (3), −0.06 (1), 0.08 (5)
J-Keep-D	−0.11 (35), −0.06 (4), 0.05 (16)
	3 Extreme FRS Diffs (Topic)
K-Keep-D	0.34 (4), 0.18 (39), −0.06 (45)
C-Keep-D	−0.14 (14), −0.07 (2), 0.00 (21)
J-Keep-D	−0.29 (44), −0.19 (29), 0.19 (35)

age precision and FRS measures. The /nostop option had little effect, even on most individual topics. The /single option had little impact on average, but some individual topics were substantially affected (we look at an example below). Disabling morphology (as per the 'None' runs) was significantly detrimental to mean average precision.

Topic K-035-T (사형제도, 여론조사 (capital punishment, survey data)): Table 10 shows that topic 35 had the largest decrease in average precison when enabling the '/single' option for Korean title queries. The stemmer produced two stemming interpretations for 사형제도 (capital punishment system). The first interpretation (used by the /single option) considered 사 (company) and 형제 (sibling) to be the stems. The alternative interpretation (discarded by the /single option) considered 사형 (capital punishment) and 제 (me) to be the stems. When /single was used, potentially useful matches were not made such as 사형이 (capital punishment), 지금까지사형 (until now capital punishment), 사형에 (in capital punishment) and 사형집행으로 (with execution).

4 Submitted Runs

Table 13 lists the mean scores of the 5 runs submitted for each language (in May 2005).

The T-01 runs were plain word-based Title-only runs for each language (with decompounding enabled, as was the case for all submitted runs which used word-based approaches). They were the same as the 'Base-T' runs of Table 3 (except that an older SearchServer build was used, but it does not appear to have made a difference).

Table 13. Mean Scores of Submitted Runs

Run	FRS	S@1	S@10	MAP
HUM-C-C-T-01	0.871	31/50	45/50	0.324
(C-Fusion-T)	0.841	32/50	44/50	0.315
HUM-C-C-T-04	0.837	26/50	44/50	0.339
HUM-C-C-D-02	0.815	24/50	43/50	0.268
HUM-C-C-D-05	0.810	22/50	43/50	0.322
HUM-C-C-DN-03	0.896	29/50	46/50	0.373
HUM-J-J-T-01	0.885	28/47	44/47	0.312
(J-Fusion-T)	0.911	28/47	45/47	0.313
HUM-J-J-T-04	0.890	26/47	45/47	0.337
HUM-J-J-D-02	0.814	17/47	41/47	0.281
HUM-J-J-D-05	0.803	21/47	42/47	0.301
HUM-J-J-DN-03	0.882	28/47	44/47	0.352
HUM-K-K-T-01	0.912	29/50	49/50	0.355
(K-Fusion-T)	0.920	33/50	48/50	0.382
HUM-K-K-T-04	0.916	31/50	49/50	0.433
HUM-K-K-D-02	0.901	33/50	46/50	0.355
HUM-K-K-D-05	0.899	35/50	46/50	0.416
HUM-K-K-DN-03	0.938	39/50	48/50	0.437

The Fusion-T runs (in parentheses in Table 13) were not actually submitted. They combined the 'Base-T' and 'Ngram-T' runs into one run by adding together the relevance scores for each document. The inclusion of n-grams increased the mean FRS and MAP scores for Japanese and Korean (compared to the T-01 run) but decreased them for Chinese.

The T-04 runs were word-based blind feedback runs in which the first 3 rows of the corresponding Fusion-T run were used to find additional query terms. Only terms appearing in at most 5% of the documents were included. Mathematically, the approach was similar to Rocchio feedback with weights of one-half for the original query and one-sixth for each of the 3 expansion rows. (The T-04 runs were the only submitted runs in which n-grams had any role.) More analysis of the feedback results is below.

The D-02 runs were plain word-based Description-only runs for each language. They were the same as the 'Base-D' runs of Table 4. For Chinese and Japanese, the mean scores were lower for the Description runs than the corresponding Title runs (T-01).

The D-05 runs were word-based blind feedback runs in which the first 3 rows of the corresponding D-02 run were used to find additional query terms (using the same expansion approach as for T-04).

The DN-03 runs used the same approach as the D-

Table 14. Mean Impact of Blind Feedback

Expt	ΔMAP (95% Conf)	vs.
K-Exp-D	0.061 (0.03, 0.09)	39-10-1
C-Exp-D	0.054 (0.03, 0.08)	40-8-2
K-Exp-T	0.052 (0.03, 0.08)	40-10-0
C-Exp-T	0.024 (0.01, 0.04)	32-17-1
J-Exp-T	0.024 (0.00, 0.05)	30-17-0
J-Exp-D	0.020 (−0.01, 0.05)	32-15-0
	ΔFRS	
K-Exp-D	−0.001 (−0.03, 0.02)	9-8-33
C-Exp-T	−0.003 (−0.03, 0.03)	7-15-28
K-Exp-T	−0.004 (−0.03, 0.02)	6-7-37
C-Exp-D	−0.004 (−0.04, 0.03)	6-12-32
J-Exp-D	−0.011 (−0.04, 0.02)	9-12-26
J-Exp-T	−0.021 (−0.04, 0.00)	4-14-29

Table 15. Per-Topic Impact of Blind Feedback

Expt	3 Extreme AP Diffs (Topic)
K-Exp-D	0.47 (9), 0.25 (16), −0.10 (13)
C-Exp-D	0.28 (23), 0.21 (16), −0.08 (39)
K-Exp-T	0.31 (9), 0.27 (13), −0.07 (4)
C-Exp-T	0.12 (38), 0.11 (36), −0.09 (28)
J-Exp-T	0.26 (18), 0.24 (36), −0.13 (31)
J-Exp-D	−0.22 (43), 0.13 (22), 0.16 (16)
	3 Extreme FRS Diffs (Topic)
K-Exp-D	0.27 (39), −0.15 (32), −0.16 (4)
C-Exp-T	0.44 (33), 0.18 (38), −0.22 (26)
K-Exp-T	−0.25 (5), −0.14 (25), 0.19 (2)
C-Exp-D	0.59 (47), 0.31 (26), −0.23 (38)
J-Exp-D	−0.25 (18), −0.21 (47), 0.14 (1)
J-Exp-T	−0.25 (41), −0.16 (44), 0.07 (47)

02 runs except that the Narrative field of the topic was additionally included in the query. (No blind feedback was used for this run.) The mean scores were higher than for the corresponding D-02 runs, but in some cases the mean scores of the T-01 runs were higher still.

4.1 Impact of Blind Feedback

Tables 14 and 15 isolate the impact of the blind feedback technique on the average precision and FRS measures. The 'Exp-T' lines subtract the 'Fusion-T' run from the 'T-04' run, and the 'Exp-D' lines subtract the 'D-02' run from the 'D-05' run. Blind feedback increased all of the mean average precision scores and decreased all of the mean FRS scores. (We found the same result for European language experiments in [12]).

In the case of Japanese titles, the impact of blind feedback was statistically significant for both mean average precision and mean FRS, but in *opposite* directions.

The blind feedback approach presumably works best if relevant documents occur at the top of the list, but from the perspective of FRS, the result was already satisfactory and adding more relevants deeper in the result list is unimportant.

If the top 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. The average precision measure may not give much weight to this damage because it may give a score close to zero either way, but the FRS measure can be substantially affected.

In practice, feedback should probably be a user-controlled technique, rather than executed blindly.

This result illustrates that the 'average precision' and 'first relevant score' measures can have very different conclusions about which technique is "significantly better" (on average) for a retrieval task.

References

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

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

[3] H. Fujii and W. B. Croft. A Comparison of Indexing Techniques for Japanese Text Retrieval. *Proceedings of SIGIR'93*, 1993.

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

[5] K. L. Kwok. Comparing Representations in Chinese Information Retrieval. *Proceedings of SIGIR'97*, 1997.

[6] J. H. Lee and J. S. Ahn. Using *n*-Grams for Korean Text Retrieval. *Proceedings of SIGIR'96*, 1996.

[7] NTCIR (NII-NACSIS Test Collection for IR Systems) Home Page. http://research.nii.ac.jp/ntcir/index-en.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] Text REtrieval Conference (TREC) Home Page. http://trec.nist.gov/.

[10] S. Tomlinson. CJK Experiments with Hummingbird SearchServer™ at NTCIR-4. *Proceedings of NTCIR-4*, 2004.

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

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

[13] Yahoo! Korea. http://kr.dic.yahoo.com/.