

# 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<sup>1</sup> 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.

## 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) |

<sup>1</sup>SearchServer™, SearchSQL™ and Intuitive Searching™ are trademarks of Hummingbird Ltd. All other copyrights, trademarks and tradenames are the property of their respective owners.

**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 Comparison 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.
- “ $\Delta$ MAP” is the difference of the mean average precision scores when subtracting the Base run from the listed run (and “ $\Delta$ FRS” is the difference of the (mean) FRS scores).
- “95% Conf” is an approximate 95% confidence interval for the mean difference (calculated from









