# CJK Experiments with Hummingbird SearchServer™ at NTCIR-4

Stephen Tomlinson
Hummingbird
Ottawa, Ontario, Canada
stephen.tomlinson@hummingbird.com
July 31, 2004

## Abstract

*Hummingbird submitted ranked result sets for the Chinese, Japanese, Korean and English Single Language Information Retrieval subtasks of the Cross-Lingual Information Retrieval Task of the 4th NII-NACSIS Test Collection for IR Systems Workshop (NTCIR-4). SearchServer's experimental option of splitting compound words (decompounding) was found to significantly increase mean average precision for Korean and modestly increase it for Japanese and Chinese. After decompounding, the differences between segmenting into words and an overlapping n-gram approach were not statistically significant for any of the 3 languages. Per-topic analysis suggested that segmentation sometimes separates proper names into unrelated shorter words while n-grams may overweight words of length greater than 'n'.* **Keywords:** *Decompounding, compound-splitting, compound-breaking, segmenting, stemming, n-grams, bigrams, robustness, Chinese (Traditional), Japanese, Korean.*

## 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 [5] and supports most of the world's major character sets and languages. The major conferences in text retrieval evaluation (NTCIR [9], CLEF [2] and TREC [11]) have provided opportunities to objectively evaluate SearchServer's support for 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-4 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-4 test collections (CLIR task) consisted of news articles from 1998 and 1999 in Chinese (Traditional), Japanese, Korean and English. Table 1 gives their sizes (the asterisks indicate that the text size includes a few duplicates that were actually not indexed; the number of documents is exact). For more details, see the CLIR overview paper [6].

**Table 1. Sizes of NTCIR-4 Document Sets**

| Language | Text Size | #Documents |
|---|---|---|
| Japanese | 815,022,439* bytes | 593,636 |
| Chinese | 555,285,156* bytes | 381,375 |
| Korean | 415,842,568 bytes | 254,438 |
| English | 692,860,409* bytes | 347,376 |

### 2.2 Indexing

SearchServer supports two approaches to indexing Asian text: segmenting into words and overlapping n-grams. In the experimental post-5.x versions of SearchServer used for this paper, the segmenter optionally split compound words (decompounding). The segmenter also performed stemming. The n-gram approach used bigrams for most Asian text. In this paper, we treat each parsing mode as a black box, but several examples are in the per-topic analysis below. The main differences from our NTCIR-3 experiments [12] are updated segmenters with the experimental decompounding option and the use of a short stopword list for each language.

### 2.3 Searching

SearchServer's 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 is the same as described last time [12]. Briefly, SearchServer dampens the term frequency and adjusts for document length in a manner similar to Okapi [10] and dampens the inverse document frequency using an approximation of the logarithm.

The use of inverse document frequency means that words that occur in fewer documents are assigned higher weight by relevance ranking. For all runs, document length importance was set to 500 by SearchServer's RELEVANCE_DLEN_IMP setting, though it wouldn't make much difference for the (mostly short) news articles used in these tests.

### 2.4 Diagnostic Runs

For the diagnostic runs listed in Tables 2 and 3, 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 used the word-based approach with decompounding enabled. (For English, decompounding was not applicable, but inflectional stemming was still performed.)

"Cmpd": Same as Base except that a different SearchServer table was used which had decompounding mode disabled.

"Ngram": Same as Base except that a different SearchServer table was used which was indexed with overlapping n-grams.

"Idf": Same as Base except that the search used RELEVANCE_METHOD 'V2:4' to square the importance of inverse document frequency to the weighting (compared to the other runs which used 'V2:3').

"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 (not this year's topics); an older list was used for English.

**Table 2. Scores of Diagnostic Title-only Runs**

| Run | AvgP | Robust@10 |
|---|---|---|
| J-Idf-T | 0.310 | 54/55 (98%) |
| J-Base-T | 0.309 | 55/55 (100%) |
| J-Ngram-T | 0.307 | 55/55 (100%) |
| J-Cmpd-T | 0.291 | 54/55 (98%) |
| K-Ngram-T | 0.402 | 56/57 (98%) |
| K-Base-T | 0.379 | 54/57 (95%) |
| K-Idf-T | 0.370 | 53/57 (93%) |
| K-Cmpd-T | 0.300 | 54/57 (95%) |
| C-Ngram-T | 0.181 | 46/59 (78%) |
| C-Base-T | 0.170 | 46/59 (78%) |
| C-Idf-T | 0.169 | 45/59 (76%) |
| C-Cmpd-T | 0.163 | 45/59 (76%) |
| E-Base-T | 0.299 | 56/58 (97%) |
| E-Idf-T | 0.290 | 54/58 (93%) |

### 2.5 Evaluation Measures

The NTCIR organizers 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 60 topics. In this paper, we just count 'highly relevant' or 'relevant' as relevant. We follow the NTCIR standard of discarding topics with fewer than 3 relevant. For more details, see [6].

The primary evaluation measure in this paper is "mean average precision" based on the first 1000 retrieved documents for each topic (denoted "AvgP" in Tables 2, 3 and 12). "Average precision" for a topic is the average of the precision after each relevant document is retrieved (using zero as the precision for relevant documents which are not retrieved). The score ranges from 0.0 (no relevants found) to 1.0 (all relevants found at the top of the list). For a set of topics, all topics are weighted equally by the mean. 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.

A more experimental measure is "robustness at 10 documents" (denoted "Robust@10") which is the percentage of topics for which at least one relevant document was returned in the first 10 rows (this was one of the measures investigated in the TREC Robust Retrieval track last year [16]). This measure hides a lot of retrieval differences (particularly in recall), but it may be an indicator of a user's impression of a method's robustness across topics.

**Table 3. Scores of Diagnostic Description-only Runs**

| Run | AvgP | Robust@10 |
|---|---|---|
| J-Idf-D | 0.294 | 53/55 (96%) |
| J-Base-D | 0.271 | 52/55 (95%) |
| J-Ngram-D | 0.261 | 54/55 (98%) |
| J-Cmpd-D | 0.259 | 52/55 (95%) |
| J-Keep-D | 0.249 | 53/55 (96%) |
| K-Idf-D | 0.349 | 53/57 (93%) |
| K-Base-D | 0.322 | 54/57 (95%) |
| K-Keep-D | 0.300 | 54/57 (95%) |
| K-Ngram-D | 0.295 | 54/57 (95%) |
| K-Cmpd-D | 0.254 | 51/57 (89%) |
| C-Idf-D | 0.153 | 44/59 (75%) |
| C-Ngram-D | 0.148 | 42/59 (71%) |
| C-Base-D | 0.145 | 45/59 (76%) |
| C-Cmpd-D | 0.140 | 44/59 (75%) |
| C-Keep-D | 0.139 | 42/59 (71%) |
| E-Idf-D | 0.268 | 51/58 (88%) |
| E-Base-D | 0.267 | 54/58 (93%) |
| E-Keep-D | 0.251 | 52/58 (90%) |

### 2.6 Statistical Significance Tables

For Tables 4 and 8, the columns are as follows:

- "Expt" is a label for the experiment. It starts with the language ('C' for Chinese, 'J' for Japanese, 'K' for Korean, 'E' for English), followed by the feature being isolated (e.g. 'Dmpd' for decompounding), followed by the topic fields used ('T' for Titles or 'D' for Descriptions).
- "Diff" is the difference of the mean average precision scores of the two runs being compared.
- "95% Conf" is an approximate 95% confidence interval for the difference calculated using Efron's bootstrap percentile method[2] [3] (using 100,000 iterations). If zero is not in the interval, the result is "statistically significant" (at the 5% level), i.e. the feature is unlikely to be of neutral impact, though if the average difference is small (e.g. $<0.020$) it may still be too minor to be considered "significant" in the magnitude sense.
- "vs." is the number of topics on which the score was higher, lower and tied (respectively) with the feature enabled. These numbers should always add to the number of topics (59 for Chinese, 55 for Japanese, 57 for Korean, 58 for English).

[2] See [13] for some comparisons of confidence intervals from the bootstrap percentile, Wilcoxon signed rank and standard error methods for both average precision and Precision@10.

**Table 4. Impact of Decompounding on Average Precision**

| Expt | Diff | 95% Conf | vs. |
|---|---|---|---|
| K-Dmpd-T | 0.078 | ( 0.042, 0.116) | 42-15-0 |
| K-Dmpd-D | 0.069 | ( 0.036, 0.105) | 44-13-0 |
| J-Dmpd-T | 0.017 | ( 0.004, 0.033) | 30-24-1 |
| J-Dmpd-D | 0.011 | ($-0.003$, 0.028) | 31-23-1 |
| C-Dmpd-T | 0.007 | ($-0.008$, 0.026) | 32-25-2 |
| C-Dmpd-D | 0.006 | ($-0.002$, 0.014) | 32-25-2 |

### 2.7 Per-Topic Tables

For tables (such as Table 5) which focus on the per-topic impacts of a particular experiment, the columns are as follows:

- "Ranks" gives the 'absolute rank' (based on the absolute value of the difference), followed by the 'signed rank' (based on the signed value of the difference, which is also followed by an 'e' for the most extreme topic in each direction). Typically the tables contain the topics with the 10 largest absolute differences in descending order.
- "Topic" gives the topic language, followed by the topic number (1-60), followed by the topic field (always T for Title in this paper).
- "Difference" gives the difference in the average precision score, followed by the score of the Base method and the score of the other method; i.e. the difference is the Base score minus the other score (before rounding the scores to 2 decimal places).
- "Rel" is the number of relevant documents for the topic. Topics with few relevants tend to be easier to analyze, but may be more subject to chance differences.

The last row of each per-topic table has the averages over all topics (not just the listed 10).

For each per-topic table, there is a followup section with analysis of at least the extreme topic in each direction (and sometimes one more).

## 3 Impact of Decompounding

Table 4 shows the difference of the 'Base' and 'Cmpd' runs (of Tables 2 and 3) for each language and topic field. In each case, mean average precision was higher with decompounding enabled. The differences passed the statistical significance test for Korean Titles, Korean Descriptions and Japanese Titles (i.e. their confidence intervals do not contain zero). The

**Table 5. Largest Impacts of Decompounding on Average Precision, Korean Titles**

| Ranks | Topic | Difference | Rel |
|---|---|---|---|
| 1/ 1e | K3T | 0.50 (0.50-0.00) | 33 |
| 2/ 2 | K14T | 0.45 (0.47-0.03) | 67 |
| 3/ 3 | K30T | 0.40 (0.64-0.25) | 84 |
| 4/ 4 | K59T | 0.33 (0.37-0.04) | 110 |
| 5/57e | K12T | −0.29 (0.38-0.67) | 4 |
| 6/ 5 | K54T | 0.29 (0.36-0.07) | 50 |
| 7/ 6 | K5T | 0.26 (0.74-0.47) | 94 |
| 8/ 7 | K34T | 0.24 (0.65-0.41) | 4 |
| 9/ 8 | K47T | 0.23 (0.39-0.16) | 41 |
| 10/ 9 | K31T | 0.21 (0.35-0.15) | 129 |
| Avg57 | K-Dmpd-T | 0.08 (0.38-0.30) | 55 |

**Table 6. Largest Impacts of Decompounding on Average Precision, Japanese Titles**

| Ranks | Topic | Difference | Rel |
|---|---|---|---|
| 1/ 1e | J42T | 0.25 (0.40-0.15) | 56 |
| 2/ 2 | J52T | 0.24 (0.28-0.04) | 179 |
| 3/ 3 | J51T | 0.12 (0.73-0.61) | 58 |
| 4/ 4 | J59T | 0.11 (0.65-0.54) | 233 |
| 5/ 5 | J34T | 0.09 (0.12-0.03) | 60 |
| 6/55e | J39T | −0.07 (0.23-0.29) | 58 |
| 7/ 6 | J28T | 0.06 (0.18-0.12) | 56 |
| 8/ 7 | J47T | 0.05 (0.31-0.26) | 128 |
| 9/ 8 | J53T | 0.04 (0.71-0.67) | 45 |
| 10/54 | J21T | −0.04 (0.83-0.87) | 16 |
| Avg55 | J-Dmpd-T | 0.02 (0.31-0.29) | 130 |

impact for Korean was particularly substantial (almost 8 points for Korean Titles).

These decompounding results are consistent with what we have seen when investigating European languages, particularly German, Finnish, Dutch and Swedish [14]. It appears that for any language with a lot of compound words, decompounding is likely to be a useful technique (on average).

### 3.1 Korean Decompounding

Topic 3: Table 5 shows that the largest impact from decompounding for Korean Title queries was on topic 3, which consisted of one compound word 배아줄기세포 (Embryonic Stem Cells). Without decompounding, the query scored zero because 배아줄기세포 (or any inflection) did not occur in the documents. Decompounding the query produced the stems 배아 (embryo bud), 줄기 (stem) and 세포 (cell). 배아 was a stem of many compound words in relevant documents such as 배아단계까지, 배아단계는, 배아단계의 (embryo bud phase), 배아복제는 and 복제배아의 (embryo bud reproduction). 세포 was also a stem of inflections and compounds in relevant documents such as 세포가 (cell), 4세포기의 (4th cell phase), 세포분열을, 세포분열이 (cell division), 세포분열과정을 (cell division process) and 제거하고체세포 (somatic cell removal). 줄기 (stem) did not commonly appear in the relevant documents. It appears that many of the related words in the documents were different compound words, so it seems unlikely any simple query could have scored highly without decompounding document words (or at least doing arbitrary word breaks using n-grams).

Topic 14: The query 환경 호르몬 (Environmental Hormone) was provided as 2 words. But the relevant documents usually used 환경호르몬 or other 1-word variants such as 환경호르몬을 or 환경호르몬의 which only matched the query when decompounding mode was used. It may also have been helpful that decompounding would allow compounds containing an 'environment' component to match, e.g. 환경전문가들은 (environment specialists), 환경보호국 (environment protecting state), or compounds containing a 'hormone' component to match, e.g. 남성호르몬 (male hormone), 여성호르몬 (female hormone). Generally, this topic illustrates that it can be difficult for a user to know whether a term is used as a compound word in the documents or not, and with a decompounding system it should not matter if the user guesses right.

Topic 12: The biggest negative impact from decompounding was on topic 12, for which decompounding mode split 구로사와 (Kurosawa, the surname of a Japanese film director) to common words 구로 (a region in Seoul) and 사 (buy, after stemming). With decompounding off, 구로사와 was stemmed to uncommon word 구로사, giving higher weight to a good term for matching Kurosawa documents, which appears to be why it scored higher.

### 3.2 Japanese Decompounding

Topic 42: Table 6 shows that the largest impact from decompounding for Japanese Title queries was on topic 42, for which the compound word アップルコンピュータ (Apple Computer) was split to アップル (Apple) and コンピュータ (Computer), and average precision was 25 points higher. The main reason appears to be that some relevant documents did not use アップルコンピュータ but just used アップル or sometimes the hyphenated form of アップル・コンピュータ. Also, sometimes relevant documents would just use the long form once and the short form afterwards; e.g. JA-981016069 is a highly relevant document which has アップルコンピュータ

just once, but アップル 5 other times, so matching the short form was helpful for the document to appear on topic. It may also have been helpful that splitting to two terms in effect doubled the weight as アップル and コンピュータ were almost as uncommon on their own as アップルコンピュータ. (The topic also contained the phrase 新製品 (New Products) which was segmented to 新 (New) and 製品 (Products) even without decompounding on.)

Topic 52: Decompounding 皇太子妃 (Crown Princess) to 皇太子 (Crown Prince) and 妃 (Princess) led to a 24 point increase. The other query term, 雅子 (Masako, the name of the princess) was left intact in both cases. The reasons that decompounding was helpful varied. For example, at least one relevant document (JY-19980616J1OYMAO1400010) used the split form 皇太子・雅子妃, which would only match in the decompounding case. Another relevant document did not contain the compound 皇太子妃 but did contain 皇太子 and 妃 separately (JY-19990114J1TYEUG0400010). Another relevant document did not contain 妃, just 皇太子 and 雅子 (JY-19991210J1TYEUG0400010). Decompounding in effect doubled the weight on 皇太子妃 because each piece was almost as uncommon on its own, making the documents with just 雅子 (the name is also used by non-princesses) rank lower. Also, a lot of relevant documents used 皇太子 repeatedly, i.e. 皇太子 was a good indicator that the document may be on topic.

Topic 39: Decompounding split 労働者 (worker) to 労働 (work) and 者 (person), and it split 外国人 (foreigner) to 外国 (foreign country) and 人 (person), which was not helpful in this case. These are words which are compounds in Japanese but not in English. The opposite also occurred in this topic in that the word 人権 (human rights) was not split, even though it corresponds to a phrase in English. Note that 外国人労働者 (foreign worker) was segmented to 外国人 (foreigner) and 労働者 (worker) even in non-decompounding mode. This topic had the largest negative impact from decompounding Japanese Titles, but it was just a 7 point drop.

### 3.3 Chinese Decompounding

Topic 7: Table 7 shows that the largest impact from decompounding for Chinese Title queries was on topic 7, for which the compound word 巴拿馬運河 (Panama Canal) was split to 巴拿馬 (Panama) and 運河 (Canal) and average precision was 38 points higher. One reason is that some relevant documents (e.g. udn_xxx_19991215_0456) referred to 巴拿馬 and 運河 but not 巴拿馬運河 together. Also, even if 巴拿馬運河 did occur, often 巴拿馬 and 運河 would also occur on their own and so would help the document to appear on topic when decompounding. 巴拿馬 and 運河 were also uncommon enough on their own for splitting them to give more combined weight from inverse document frequency than 巴拿馬運河 on its own (there were some other terms in the query).

**Table 7. Largest Impacts of Decompounding on Average Precision, Chinese Titles**

| Ranks | Topic | Difference | Rel |
|---|---|---|---|
| 1/ 1e | C7T | 0.38 (0.56-0.18) | 7 |
| 2/59e | C22T | −0.20 (0.13-0.33) | 4 |
| 3/ 2 | C52T | 0.14 (0.36-0.22) | 3 |
| 4/ 3 | C51T | 0.10 (0.42-0.32) | 13 |
| 5/58 | C48T | −0.07 (0.31-0.38) | 17 |
| 6/ 4 | C45T | 0.06 (0.13-0.07) | 47 |
| 7/57 | C5T | −0.06 (0.41-0.47) | 7 |
| 8/ 5 | C21T | 0.04 (0.31-0.27) | 17 |
| 9/ 6 | C11T | 0.04 (0.36-0.32) | 27 |
| 10/56 | C38T | −0.03 (0.33-0.37) | 5 |
| Avg59 | C-Dmpd-T | 0.01 (0.17-0.16) | 22 |

Topic 22: The next largest impact was on topic 22, for which the company name 起亞汽車 (Kia Motors) was segmented to 起 (gets up), 亞 (Asia) and 汽車 (automobile), regardless of whether decompounding was enabled. The first two words (起 and 亞) were particularly common and matched many words unrelated to Kia. It appears the score was lower with decompounding enabled because there would be even more spurious matches in non-relevant documents, e.g. in non-relevant document cts_pol_19971203_0001, unrelated word 亞琳 was matched just when decompounding. In a manual system, a user could specify phrasing to compensate for over-segmentation.

Topic 52: Unlike for Japanese (discussed earlier), initial segmenting of 皇太子妃 (Crown Princess) produced 皇太子 (Crown Prince) and 妃 (Princess), and decompounding further split 皇太子 to 皇 (Emperor) and 太子 (shorter form of Crown Prince). One relevant document (cts_int_19981210_0010) was not matched by the longer form of Crown Prince; it just contained the shorter form in 日太子妃雅子生日 (Japanese Crown Princess Masako's birthday), so the decompounding mode scored higher.

## 4 Comparison with N-grams

Table 8 shows the difference of the 'Base' and 'Ngram' runs (of Tables 2 and 3), i.e. decompounded words vs. overlapping n-grams for each language and topic field. A positive difference means the word-based approach scored higher. None of the differences passed the statistical significance test (i.e. their confidence intervals all contain zero). Per-topic analysis for the Title topics follows.

**Table 8. Differences of Words vs. N-grams in Average Precision**

| Expt | Diff | 95% Conf | vs. |
|---|---|---|---|
| J-Seg-T | 0.002 | (−0.016, 0.018) | 29-25-1 |
| C-Seg-T | −0.010 | (−0.034, 0.010) | 31-27-1 |
| K-Seg-T | −0.023 | (−0.055, 0.005) | 25-32-0 |
| K-Seg-D | 0.027 | (−0.002, 0.056) | 37-20-0 |
| J-Seg-D | 0.010 | (−0.011, 0.032) | 30-24-1 |
| C-Seg-D | −0.002 | (−0.024, 0.017) | 35-24-0 |

**Table 9. Largest Differences of Words vs. N-grams in Average Precision, Japanese Titles**

| Ranks | Topic | Difference | Rel |
|---|---|---|---|
| 1/55e | J26T | −0.25 (0.17-0.42) | 63 |
| 2/54 | J10T | −0.19 (0.17-0.36) | 55 |
| 3/ 1e | J50T | 0.14 (0.62-0.48) | 299 |
| 4/ 2 | J24T | 0.12 (0.22-0.10) | 83 |
| 5/ 3 | J51T | 0.11 (0.73-0.62) | 58 |
| 6/53 | J5T | −0.11 (0.18-0.29) | 38 |
| 7/ 4 | J12T | 0.10 (0.13-0.03) | 52 |
| 8/ 5 | J7T | 0.09 (0.46-0.37) | 12 |
| 9/52 | J39T | −0.08 (0.23-0.31) | 58 |
| 10/ 6 | J43T | 0.08 (0.59-0.50) | 173 |
| Avg55 | J-Seg-T | 0.00 (0.31-0.31) | 130 |

### 4.1 Japanese Words vs. N-grams

Topic 26: Table 9 shows that the largest difference of words vs. n-grams for Japanese Title queries was on topic 26. The only difference in the parsing of the query was for the phrase 外交関係 (Diplomatic Relations), for which the segmenter (even if decompounding was not enabled) produced two words: 外交 (Diplomacy) and 関係 (Relationship), while the n-gram approach additionally produced the middle bigram 交関 which was less common in the documents, in effect giving the phrase higher weight, which turned out to be helpful for this query.

Topic 50: The query term 地下核実験 (underground nuclear testing) was segmented to 地下 (underground), 核 (nuclear) and 実験 (experiments). The n-gram method would not match all occurrences of the 1-character word 核 (just occurrences matching bigrams 下核 or 核実), so it would not in general match related text such as 核拡散防止条約 (Nuclear Non-Proliferation Treaty), 核兵器 (nuclear weapons) or 核開発 (nuclear development), which may be why n-grams scored lower on this topic.

**Table 10. Largest Differences of Words vs. N-grams in Average Precision, Chinese Titles**

| Ranks | Topic | Difference | Rel |
|---|---|---|---|
| 1/59e | C22T | −0.49 (0.13-0.62) | 4 |
| 2/58 | C3T | −0.18 (0.26-0.44) | 16 |
| 3/ 1e | C7T | 0.17 (0.56-0.39) | 7 |
| 4/ 2 | C60T | 0.15 (0.32-0.16) | 22 |
| 5/57 | C46T | −0.15 (0.16-0.31) | 13 |
| 6/56 | C5T | −0.14 (0.41-0.55) | 7 |
| 7/55 | C48T | −0.09 (0.31-0.41) | 17 |
| 8/54 | C59T | −0.09 (0.02-0.11) | 19 |
| 9/53 | C36T | −0.09 (0.17-0.25) | 19 |
| 10/ 3 | C2T | 0.08 (0.31-0.22) | 17 |
| Avg59 | C-Seg-T | −0.01 (0.17-0.18) | 22 |

### 4.2 Chinese Words vs. N-grams

Topic 22: Table 10 shows that the largest difference of words vs. n-grams for Chinese Title queries was on topic 22. We've already seen that decompounding was harmful for this topic, but even with it disabled, n-grams would have scored 29 points higher (0.62-0.33). N-grams gave much higher weight to 起亞汽車 (Kia Motors) by producing uncommon bigrams 起亞 (Kia) and 亞汽 (which emphasized the phrase). Recall that the segmenter produced common 1-character words 起 and 亞 (and assigned them low weights because they were common). Both approaches produced 汽車 (Motors) with similar weight. Although the segmenter still matched 起亞汽車 in the documents, its weighting favored non-relevant documents with other query terms before retrieving the Kia Motors documents.

Topic 7: The n-gram weighting in effect doubled the weight on the odd-lengthed word 巴拿馬 (Panama, parsed as 2 bigrams: 巴拿 and 拿馬) compared to Canal (運河), apparently boosting the rank of non-relevant documents which focused on Panama (such as udn_xxx_19991008_0423 which was ranked 10th using n-grams but a lower 17th using the segmenter). (The topic also contained other words which had parsing differences, but the Panama Canal differences seemed most important.)

### 4.3 Korean Words vs. N-grams

Topic 2: Table 11 shows the largest difference of word vs. n-grams for Korean Title queries was on topic 2. The query contained 죠니워커 which appears to be a misspelling of 조니워커 (Johnnie Walker). (The Description and the documents used the correct form, but the Title and Concepts did not. The first character 죠 sounds like 'Jyo' while 조 sounds like 'Jo'.) The

**Table 11. Largest Differences of Words vs. N-grams in Average Precision, Korean Titles**

| Ranks | Topic | Difference | Rel |
|---|---|---|---|
| 1/57e | K2T | −0.51 (0.02-0.53) | 13 |
| 2/56 | K12T | −0.47 (0.38-0.85) | 4 |
| 3/ 1e | K52T | 0.22 (0.63-0.41) | 3 |
| 4/55 | K50T | −0.21 (0.05-0.25) | 36 |
| 5/54 | K40T | −0.20 (0.19-0.39) | 18 |
| 6/53 | K54T | −0.18 (0.36-0.54) | 50 |
| 7/ 2 | K59T | 0.16 (0.37-0.20) | 110 |
| 8/ 3 | K21T | 0.14 (0.35-0.21) | 22 |
| 9/52 | K38T | −0.09 (0.16-0.25) | 32 |
| 10/ 4 | K42T | 0.09 (0.61-0.52) | 27 |
| Avg57 | K-Seg-T | −0.02 (0.38-0.40) | 55 |

query form was not found in the documents when using the segmenter. The n-gram approach, however, included the bigrams 니워 and 워커 with high weight and scored much higher. The bigram 죠니 (Jyo-ni) was produced with even higher weight (because it was very uncommon in the documents), but it didn't appear often enough with other query terms to bring in a lot of non-relevant documents. A manual system might suggest to the user a modified query with the correct spelling.

Topic 12: In topic 12, the Kurosawa topic mentioned earlier, the query word 아키타 (Akita) appears to be a misspelling of 아키라 (Akira). The segmenter gave a high weight to the 3-character query form (and also a low weight to the split Kurosawa as mentioned earlier) and so preferred some non-relevant documents containing 아키타, while the n-gram approach produced enough useful bigrams to not be too distracted by the misleading bigram 키타.

Topic 52: Segmenting in decompounding mode actually left the query word 황태자비 (Crown Princess) intact and the 3 relevant documents were found with high precision. The n-gram approach produced overlapping bigrams and matches such as 황태자, 태자 and 마의태자는 lowered its precision.

## 5 Submitted Runs

Table 12 lists the scores of the 5 runs submitted for each language (in November 2003). All of the submitted runs used the word-based approach with decompounding enabled.

The T-01 runs were plain Title-only runs for each language. They were the same as the 'Idf-T' runs of Table 2 except that a different experimental version of SearchServer was used (including an older version of the segmenter).

**Table 12. Scores of Submitted Runs**

| Run | AvgP | Robust@10 |
|---|---|---|
| HUM-J-J-T-01 | 0.311 | 54/55 (98%) |
| HUM-J-J-D-02 | 0.289 | 53/55 (96%) |
| HUM-J-J-C-03 | 0.274 | 54/55 (98%) |
| HUM-J-J-T-04 | 0.340 | 53/55 (96%) |
| HUM-J-J-D-05 | 0.317 | 54/55 (98%) |
| HUM-K-K-T-01 | 0.366 | 52/57 (91%) |
| HUM-K-K-D-02 | 0.334 | 53/57 (93%) |
| HUM-K-K-C-03 | 0.348 | 53/57 (93%) |
| HUM-K-K-T-04 | 0.401 | 52/57 (91%) |
| HUM-K-K-D-05 | 0.368 | 50/57 (88%) |
| HUM-C-C-T-01 | 0.169 | 45/59 (76%) |
| HUM-C-C-D-02 | 0.157 | 45/59 (76%) |
| HUM-C-C-C-03 | 0.170 | 46/59 (78%) |
| HUM-C-C-T-04 | 0.184 | 44/59 (75%) |
| HUM-C-C-D-05 | 0.179 | 46/59 (78%) |
| HUM-E-E-T-01 | 0.290 | 54/58 (93%) |
| HUM-E-E-D-02 | 0.266 | 51/58 (88%) |
| HUM-E-E-C-03 | 0.286 | 55/58 (95%) |
| HUM-E-E-T-04 | 0.310 | 53/58 (91%) |
| HUM-E-E-D-05 | 0.299 | 50/58 (86%) |

The D-02 runs were the same as the T-01 runs except that the Descriptions were used instead of the Titles. The Title runs scored higher in mean average precision than the Description runs for each language (and the differences were statistically significant for Korean and Japanese).

The C-03 runs were the same as the T-01 runs except that the Concepts were used instead of the Titles. (The Concepts were usually longer keyword lists than the Titles.) The Titles scored higher than the Concepts (on average) except in Chinese, but these differences were not statistically significant.

The T-04 and D-05 runs were blind feedback runs in which the first 3 rows of the corresponding plain run were used to find additional query terms. Only terms appearing in at most 5% of the documents were included. Mathematically, the approach is similar to Rocchio feedback with weights of one-half for the original query and one-sixth for each of the 3 expansion rows. The results were similar to what we saw for English [15] in the recent TREC Robust Retrieval track, i.e. the increases in mean average precision were statistically significant, but the Robust@10 score was often lower.

As usually happens in ad hoc search evaluations, most groups' runs (including ours) had similar mean average precision scores with a bunching at the top (the median scores were close to the highest scores). When comparing our submitted runs to the other groups by topic, the ones scoring below median were

generally the topics for which Tables 9, 10 and 11 showed that n-grams scored higher (we did not use n-grams in any of our submitted runs, just in diagnostic runs). Our blind feedback technique (just used in some submitted runs, not in diagnostic runs) may have been less effective in this evaluation from our not bothering to discard the text in the NTCIR documents which was not considered part of the content (such as document identifiers).

## 6 Robustness Across Topics

For the experiments investigated in this paper (particularly decompounding and words vs. n-grams), the differences in the Robust@10 score were not significant (usually the difference needs to be at least 4 topics (sometimes more) to pass the statistical significance test when there are 55-59 topics in total). Even for the Korean Title decompounding case, there was no net change in robustness (3 topics were gained, but 3 others were lost).

## 7 Related Work and Conclusions

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. [7] prefers "short-words" to "long-words" for Chinese, [4] prefers "short unit keywords" to "long unit keywords" for Japanese, and [8] finds that "morphemes" or "simple nouns" are more effective than "words" or "compound nouns" for Korean). In this paper, we measured the impact of the experimental "decompounding" option of our implementation on the NTCIR-4 CLIR ad hoc search tasks. The gains in mean average precison were particularly substantial for Korean. The per-topic analysis suggested that the gains were smaller for Japanese and Chinese in part because even in non-decompounding mode the segmenter usually produced short words.

Researchers have found that n-gram methods generally score comparably to the highest-scoring word-based approaches in CJK ad hoc search experiments [7, 4, 8] (though n-grams produce a larger index and have more search-time overhead). We found a similar result in our implementation: for the ad hoc search tasks, the differences of n-grams and (decompounded) words in mean average precision were not statistically significant for any of the 3 languages. Note, however, that n-gram methods may not be suitable for features (not included in these experiments) such as spelling-correction or supporting domain-specific thesauri.

The NTCIR test collections are useful for understanding the differences between retrieval methods for East Asian languages. This paper focused on experiments with automatic methods for ad hoc search. The results suggested that the segmenters work well and that decompounding mode is the better default, especially for Korean. But in general, the methods may have impacts on the user experience not covered in the described experiments, so one must be cautious when interpreting the results.

## 8 Acknowledgements

For the translations to English, the official topic translations and online services [1] were helpful, and also invaluable assistance was provided by Yiming Hu and Bryan Yoo. Two anonymous reviewers provided general feedback on the workshop paper.

## 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] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

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

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

[6] K. Kishida, Kuang-hua Chen, S. Lee, K. Kuriyama, N. Kando, Hsin-Hsi Chen, S. H. Myaeng and K. Eguchi. Overview of CLIR Task at the Fourth NTCIR Workshop. *Proceedings of NTCIR-4*, 2004.

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

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

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

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

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

[12] S. Tomlinson. Asian Language Parsing Evaluated by Hummingbird SearchServer™ at NTCIR-3. *Proceedings of NTCIR-3*, 2003.

[13] S. Tomlinson. Experiments in 8 European Languages with Hummingbird SearchServer™ at CLEF 2002. *Working Notes for the CLEF 2002 Workshop*, 2002.

[14] S. 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] S. Tomlinson. Robust, Web and Genomic Retrieval with Hummingbird SearchServer™ at TREC 2003. *Proceedings of TREC 2003*, 2004.

[16] E. M. Voorhees. Overview of the TREC 2003 Robust Retrieval Track. *Proceedings of TREC 2003*, 2004.