Monolingual Experiments with Far-East Languages in NTCIR-6

Samir ABDOU, Jacques SAVOY
Computer Science Dept, University of Neuchatel
rue Emile Argand 11, 2009 Neuchatel, Switzerland
{ Samir.Abdou, Jacques.Savoy }@unine.ch

Abstract

This paper describes our third participation in an evaluation campaign involving the Chinese, Japanese and Korean languages (NTCIR-6). Our participation is motivated by three objectives: 1) study the retrieval performances of various probabilistic and language models for these languages; 2) compare the relative retrieval effectiveness of a combined "unigram & bigram" indexing scheme combined with an automatic word-segmenting approach for Chinese and Japanese languages; and 3) evaluate the relative performance of the various data fusion strategies used to combine separate result lists in order to enhance retrieval effectiveness.

Keywords: *CLIR, Chinese, Japanese and Korean languages, probabilistic IR model, language model, evaluation.*

1 Overview of NTCIR-6 Test Collection

The sixth NTCIR evaluation campaign is divided in two stages. During the first, we used the test collections built during the NTCIR-5 campaign (Table 1 shows several related statistics and more details can be found in [9]). As requests the organizers reused the topic descriptions developed during the NTCIR-3 and NTCIR-4 campaigns. These topics were originally created for document collections covering news from various years (1998-99, and 1994 for covering Korean topics). Also used were NTCIR-5 document collections extracted from newspapers published during the years 2000-01. Clearly, due to this time difference, not all NTCIR-3 and NTCIR-4 topics would produce relevant answers. In this first stage a subset of 50 requests was thus selected by the organizers to form the topic descriptions. For each selected topic, relevant items can be found in the three languages.

In this paper, when analyzing the number of pertinent documents per topic, we only considered rigid assessments and thus only "highly relevant" and "relevant" items were seen as being relevant. A comparison of the number of relevant documents per topic, as shown in Table 1, indicates that for the Chinese collection the median number of relevant items per topic is 41.5, a value similar to that of the Japanese corpus (43), while for the Korean collection it was only 24.5. Clearly, the number of relevant articles was greater for the Japanese (3,180) corpus, when compared to the Chinese (2,598), or Korean (2,280) collections.

Stage 1	Chinese	Japanese	Korean
Document	NTCIR-5	NTCIR-5	NTCIR-5
# documents	901,446	858,400	220,374
year	2000-01	2000-01	2000-01
Query from	NTCIR-3 & 4	NTCIR-3 & 4	NTCIR-3 & 4
# of queries	**50**	**50**	**50**
# of rel. items	2,598	3,180	2,280
mean	51.96	63.6	45.6
median	41.5	43	24.5

Table 1. Various statistics concerning Stage1

Following the TREC model, the structure of each topic consisted of four logical sections: a brief title ("<TITLE>" or T), a one-sentence description ("<DESC>" or D), a narrative part ("<NARR>" or N) specifying both the topic's background context ("<BACK>") and relevance assessment criteria ("<REL>"), and finally a concept section ("<CONC>" or C) providing a few related terms. Rather than limiting available topics to a narrow subject range, those chosen reflect a variety of information needs (such as "Teenager's Fashion" (Query #110), "International incidents at Sea" (Query #19), "TV Programs on New Year Holidays" (Query #100), "Computer virus", (Query #74) or "North Korea, Starvation, Response" (Query #37)).

During the various NTCIR campaigns, the mandatory runs are based on either the title-only (or T) or on the description-only (or D) sections of the topic descriptions. In our evaluations and in order to improve relevance assessments, we also conducted experiments using both the topic description and narrative sections (DN).

In the second stage of the NTCIR-6 campaign, we wanted to verify and measure retrieval performance consistency across the three test-collections. To do so the organizers reused the NTCIR-3 to NTCIR-5 corpora (both documents and topic descriptions). As shown in Table 2, the queries were used to search

different document collections. For the Japanese corpus for example, queries from NTCIR-3 had to be searched against the articles extracted from the *Mainichi* (1998-99, 298 MB), while NTCIR-4 queries had to search for responses using both the *Mainichi* (1998-99) and the *Yomiuri* (1998-99) newspapers (a total of 733 MB). Finally, the NTCIR-5 topics had to be evaluated using the *Mainichi* (2000-01) and the *Yomiuri* (2000-01) newspapers (for a total of 1,100 MB). As shown in Table 2, for NTCIR-3 and 4 the Chinese collection is the same (for more information, see [9]). Finally, in order to obtain really comparable results, the IR models had to be "frozen" when searching across NTCIR-3 through NTCIR-5 corpora, meaning that during this second stage the values of each parameter had to be fixed for all searches (there was no fine tuning according to the specific document collection).

Stage 2	Chinese	Japanese	Korean
NTCIR-3			
size	490 MB	298 MB	68 MB
# docum.	381,681	220,078	66,146
year	1998-99	1998-99	1994
# of queries	**42**	**42**	**30**
# rel. items	1,928	1,654	2,081
mean	45.9	39.4	69.4
median	26	18	35.5
NTCIR-4			
size	490 MB	733 MB	370 MB
# docum.	381,681	596,058	254,438
year	1998-99	1998-99	1998-99
# of queries	**59**	**55**	**57**
# rel. items	1,318	7,137	3,131
mean	22.3	129.8	54.9
median			
NTCIR-5			
size	1,100 MB	1,100 MB	312 MB
# docum.	901,446	858,400	220,374
year	2000-01	2000-01	2000-01
# of queries	**50**	**47**	**50**
# rel. items	1,885	2,112	1,829
mean	37.7	44.9	36.6
median	26	24	25.5

Table 2. Various statistics from Stage 2

2 Indexing and Searching Strategies

In order to draw useful conclusions when analyzing test-collections, we considered it important to evaluate the retrieval performance using the best-performing IR models, namely both the probabilistic and language models paradigms

To achieve this we implemented the well-known Okapi model (or BM25) [12]. The probabilistic family of models is not however limited to the Okapi approach, and thus we also implemented approaches based on the *Divergence from Randomness* (DFR) framework [3], making use of two information measures. These included Inf^1 (measuring informative content of the document as compared to the entire collection), and Inf^2 (measuring information gain with respect to the *elite* set, the set of documents in which the underlying term occurs). To reflect the indexing weight w_{ij} attached to term t_j in document D_i, we have:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}] \cdot (1 - Prob^2_{ij}) \quad (1)$$

in which $Prob^1_{ij}$ is the probability of having by pure chance tf_{ij} occurrences of the term t_j in a document (and various probabilistic models could be used to estimate this probability). On the other hand, $Prob^2_{ij}$ is the probability of encountering a new occurrence of term t_j in the given document, once tf_{ij} occurrences of this term have already been found. There are various distributions or probabilistic laws that can be used in models to obtain a quantitative evaluation of this framework.

As a first model, we implemented the PB2 scheme, defined by the following equations:

$$Inf^1_{ij} = -\log_2[(e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}})/tf_{ij}!] \quad (2)$$

$$Prob^2_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij} + 1))] \quad (3)$$

$$\text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl) / l_i)]$$

$$\text{and } \lambda_j = tc_j / n$$

where tc_j indicates the number of occurrences of term t_j in the collection, l_i the length (number of indexing terms) of document D_i, *mean dl* the average document length, n the number of documents in the corpus, and c a constant.

As a second variant, the model I(n)B2 is based on another evaluation for Inf^1 component, defined as follows:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (df_j + 0.5)] \quad (4)$$

where df_j indicates the number of documents indexed using the term t_j. To evaluate $Prob^2$, we still apply Equation 3.

As a third variant, always within the DFR framework, we used the IFB2 model, defined as follows:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (tc_j + 0.5)] \quad (5)$$

Finally, we also considered an approach based on a language model (LM) [7], [8], known as a non-parametric probabilistic model (the Okapi and DFR are viewed as parametric models). Probability estimates would thus not be based on any known distribution (as in Equation 2, 3, 4 or 5) but rather be estimated directly, based on occurrence frequencies in document D or corpus C. Within this language model paradigm, various implementations and smoothing methods might be considered. In this study for example we adopted a model proposed by Hiemstra [8], as described in Equation 6, which combines an estimate based on document (denoted by $P[t_j | D_i]$) and on corpus (represented by $P[t_j | C]$).

$$P[D_i | Q] = P[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot P[t_j | D_i] + (1-\lambda_j) \cdot P[t_j | C]] \quad (6)$$

with $P[t_j \mid D_i] = tf_{ij}/l_i$
and $P[t_j \mid C] = df_j/lc$ with $lc = \sum_k df_k$

where λ_j is a smoothing factor (constant for all indexing terms t_j, and usually fixed at 0.35) and *lc* reflecting the size of the corpus C.

In defining these probabilistic and language models, we implicitly admitted that words are our indexing unit. To achieve this for the Japanese language, each sentence was automatically segmented using the morphological analyzer ChaSen [11], and for the Chinese corpus each was segmented using Mandarin Tools (freely available at www.mandarintools.com).

In the Korean language, words are clearly delimited and thus automatic segmentation was not required. In Korean however it is known that compound constructions frequently exist, and that they could harm retrieval performance. Thus we applied the *Hangul Analyser Module* tool (HAM, nlp.kookmin.ac.kr) in order to automatically decompose them.

In addition to these word-based indexing strategies, we also indexed documents by applying a combined "unigram & bigram" indexing scheme. To generate the corresponding bigrams, we used an overlapping bigram approach, an indexing scheme found to be effective for various Chinese collections [10], or during previous NTCIR campaigns [5], [14], [1], [2]. Based on this technique for example, the sequence "ABCD EFG" would generate the following bigrams {"AB," "BC," "CD," "EF," and "FG"}. In our work, we generated these overlapping bigrams for Asian ideograms only, using Latin characters, digits, spaces and other punctuation marks (collected for each language in their respective encoding) in order to stop bigram generation. Moreover, we did not split any words written in ASCII characters.

For the Korean collection, we only considered the bigram approach, given that it tended to result in the best MAP in addition to the word-based indexing strategy [2]. For the Chinese and Japanese languages, previous experiments [2] tended to demonstrate that combining both unigrams (or characters) and bigrams, when indexing documents and queries tended to produce better MAP than did simple bigrams. Based on this, we only considered this combined indexing strategy for the Chinese and Japanese languages.

Of course not all unigrams and bigrams are always useful for retrieving pertinent answers, thus the most frequent terms might be be removed before indexing. For the Chinese language, we defined a list of the 39 most frequent unigrams, the 49 most frequent bigrams plus a list of 91 words (used when applying a word-based indexing scheme in Chinese). For Japanese we defined a short stopword list of 30 words and another of 20 bigrams, and for Korean our stoplist was composed of 91 bigrams.

Before generating the bigrams for the Japanese documents, we removed all Hiragana characters, given that these characters are used for grammatical purposes to write words (e.g., *doing*, *in*, *of*), as well as the inflectional endings for verbs, adjectives and nouns. Moreover, half-width characters were replaced by their corresponding full-width version.

3 Evaluation of Various IR Models

To measure retrieval performance, we adopted the mean average precision (MAP) computed by the `trec_eval` package. In the following tables the best performance under a given condition is shown in bold. MAP values obtained by the different probabilistic models applying three different query formulations (T, D, DN) are reported in Tables 3 to 5 for Stage 1 (for the Chinese, Japanese and Korean languages respectively). For Stage 2, the evaluations are shown in Tables 6 to 8, corresponding to the NTCIR-3 through NTCIR5 test-collections.

	MAP – **Chinese** (Stage 1, 50 queries)		
	T	**D**	**DN**
PB2 unibi	0.2145	0.2084	0.2595
I(*n*)B2 unibi	**0.2157**	**0.2118**	**0.2709**
Okapi unibi	0.2109	0.1932	0.2625
LM2a MTseg	0.1673	0.1544	0.2284
PB2c2 MTseg	0.2073	0.2066	0.2459
I(*n*)B2 MTseg	0.2048	0.2003	0.2511
Okapi MTseg	0.2104	0.1970	0.2454

Table 3. MAP of various IR models (using "unigram & bigram" or MT segmentation)

	MAP - **Japanese** (Stage 1, 50 queries)		
	T	**D**	**DN**
PB2c7c4 unibi	0.2359	0.2302	0.2613
IFB2 unibi	0.2120	0.2097	0.2526
Okapid4 unibi	**0.2432**	**0.2366**	**0.2734**
PB2c3d5 CHA	0.2315	0.2160	0.2661
IFB2 c5 CHA	0.2258	0.2052	0.2630
Okapi d4 CH	0.2293	0.2125	0.2635

Table 4. MAP of various IR models (using "unigram & bigram" or Chasen segmentation)

	MAP - **Korean** (Stage 1, 50 queries)		
	T	**D**	**DN**
LM2a bigram	0.3538	0.3297	0.3891
PB2c2d0	**0.3570**	**0.3672**	**0.4242**
I(*n*)B2 bigram	0.3392	0.3387	0.3975
Okapid55	0.3492	0.3341	0.4106
LM2a HAM	0.3069	0.3058	0.3868
PB2c2d0 HAM	0.3074	0.3276	0.4008
I(*n*)B2 HAM	0.2947	0.3073	0.3893
Okapid55 HAM	0.3031	0.3106	0.4047

Table 5. MAP of various IR models (Korean corpus, using bigram or HAM decomposing)

After inspecting this data, we were able to come to the following general conclusions. For the

Chinese language, the combined "unigram & bigram" indexing strategy usually resulted in better IR performance, when compared to the word-based approach (automatic segmentation done by Mandarin Tools). For the Japanese language, the language-independent "unigram & bigram" indexing scheme usually resulted in better retrieval performance than did the corresponding word-based approach (segmentation done using the Chasen module). For the Korean language, the simple bigram indexing strategy produced better MAP values than did the decompounding scheme (HAM module).

	MAP – **Chinese** (Stage 2)		
NTCIR-3 / 42	**T**	**D**	**DN**
PB2 unibi	0.2276	0.2241	0.2693
I(*n*)B2 unibi	0.2339	**0.2303**	0.2797
LM unibi	0.2228	0.1998	0.2814
Okapi unibi	**0.2361**	0.2229	**0.2828**
LM MTseg	0.1948	0.1818	0.2587
PB2 MTseg	0.2076	0.2167	0.2617
I(*n*)B2 MTseg	0.2049	0.2080	0.2650
Okapi MTseg	0.2133	0.2002	0.2649
NTCIR-4 / 59	**T**	**D**	**DN**
PB2 unibi	0.2005	0.1885	**0.2556**
I(*n*)B2 unibi	0.1983	0.1849	0.2439
LM2 unibi	0.1852	0.1664	0.2383
Okapi unibi	0.1934	0.1727	0.2396
LM MTseg	0.1728	0.1583	0.2341
PB2 MTseg	**0.2009**	**0.1890**	0.2532
I(*n*)B2 MTseg	0.1932	0.1838	0.2479
Okapi MTseg	0.1953	0.1799	0.2439
NTCIR-5 / 50	**T**	**D**	**DN**
PB2 unibi	**0.3433**	0.3183	0.4214
I(*n*)B2 unibi	0.3404	**0.3215**	**0.4245**
Okapi unibi	0.3321	0.2892	0.4112
LM MTseg	0.2800	0.2509	0.3948
PB2 MTseg	0.3246	0.2974	0.4136
I(*n*)B2 MTseg	0.3247	0.3023	0.4206
Okapi MTseg	0.3230	0.2816	0.4135

Table 6. MAP of various IR models (using unigram & bigram or MT segmentation)

Our performance analyzes across the various probabilistic models for both NTCIR-6 stages revealed that some models were usually more effective with a given language (or test-collections). For the Chinese language (Tables 3 and 6), the I(*n*)B2 model tended to perform best (in fact 6 times over 12 evaluations) and the PB2 approach ranked in second position (4 times). For the Japanese language evaluations (see Table 4 and 7), the Okapi model performed best 8 times out of 12. The NTCIR-3 collection (top part of Table 7) was an exception to this rule.

For the Korean language the PB2 model performed best (its MAP ranked best 10 times). For the Japanese corpus the NTCIR-3 collection (top part of Table 8) there were two exceptions to this rule, while for the Korean language there were only 30 queries. It came as a surprise to see that the language model (LM) was never the best performing scheme in our various evaluations.

	MAP – **Japanese** (Stage 2)		
NTCIR-3 / 42	**T**	**D**	**DN**
PB2 unibi	**0.3325**	0.3272	0.3621
IFB2 unibi	0.2651	0.2629	0.3138
Okapid4 ubi	0.3313	0.3314	0.3731
PB2c1d5 CHA	0.3228	**0.3417**	0.3732
IFB2 c5 CHA	0.3236	0.3205	**0.3818**
Okapi d4 CHA	0.3194	0.3130	0.3754
NTCIR-4 / 55	**T**	**D**	**DN**
PB2c7c4 ubi	0.3029	0.2949	0.3403
IFB2 unibi	0.2762	0.2731	0.3357
Okapid4 ubi	**0.3110**	**0.3044**	**0.3588**
PB2c3d5 CHA	0.3035	0.2943	0.3368
IFB2 c5 CHA	0.2958	0.2806	0.3404
Okapi d4 CHA	0.2883	0.2836	0.3406
NTCIR-5 / 47	**T**	**D**	**DN**
PB2c7c4 ubi	0.3037	0.2740	0.3857
IFB2 unibi	0.2741	0.2693	0.3783
Okapid4 ubi	0.3046	**0.2908**	**0.4100**
PB2c3d5 CHA	**0.3081**	0.2895	0.4001
IFB2 c5 CHA	0.2871	0.2843	0.4003
Okapi d4 CHA	0.2752	0.2900	0.4041

Table 7. MAP of various IR models (using unigram & bigram or Chasen segmentation)

	MAP – **Korean** (Stage 2)		
NTCIR-3 / 30	**T**	**D**	**DN**
LM2a bigram	**0.2788**	0.2285	0.3126
PB2c2d0 big	0.2729	0.2544	0.3420
I(*n*)B2 bigram	0.2679	0.2491	0.3462
Okapid55 big	0.2679	0.2267	0.3400
LM2a HAM	0.2508	0.2293	0.3054
PB2c2d0 HAM	0.2682	**0.2680**	0.3417
I(*n*)B2 HAM	0.2646	0.2792	**0.3529**
Okapid55 HAM	0.2468	0.2278	0.3269
NTCIR-4 / 57	**T**	**D**	**DN**
LM2a bigram	0.4223	0.3945	0.4545
PB2c2d0 big	**0.4346**	**0.4217**	**0.5005**
I(*n*)B2 bigram	0.4180	0.3940	0.4669
Okapid55 big	0.4225	0.3927	0.4800
LM2a HAM	0.3736	0.3557	0.4166
PB2c2d0 HAM	0.3999	0.3922	0.4476
I(*n*)B2 HAM	0.3858	0.3710	0.4306
Okapid55 HAM	0.3813	0.3612	0.4415
NTCIR-5 / 50	**T**	**D**	**DN**
LM2a bigram	0.3927	0.3868	0.4869
PB2c2d0 big	**0.3939**	**0.4281**	**0.5107**
I(*n*)B2 bigram	0.3917	0.4189	0.4932
Okapid55 big	0.3833	0.3964	0.4999
LM2a HAM	0.3601	0.3413	0.4580
PB2c2d0 HAM	0.3584	0.3820	0.4729
I(*n*)B2 HAM	0.3559	0.3665	0.4580
Okapid55 HAM	0.3526	0.3441	0.4621

Table 8. MAP of various IR models (Korean corpus, using bigram or HAM decomposition)

These general trends must however be interpreted with caution, due to the fairly small performance differences between the two IR models. For the Japanese language for example and the NTCIR-5 corpus (see bottom part of Table 7), the MAP differences between the word-based and "uni-bigram" indexing schemes were relatively small when using the Okapi model (for T: 0.3046 vs. 0.3011; for D: 0.2908 vs. 0.2900; for DN: 0.4086 vs. 0.4093).

4 Blind-Query Expansion

It was observed that pseudo-relevance feedback technique (blind-query expansion) seemed to be a useful in enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach [4] whereby the system was allowed to add *m* terms extracted from the *k* best ranked documents from the original query, using the following formula:

$$Q' = \alpha \cdot Q + (\beta \cdot 1/k) \cdot \sum_{j=1}^{k} w_{ij} \quad (7)$$

in which Q' denotes the new query built for the previous query Q, and w_{ij} the indexing term weight attached to the term t_j in the document D_i. In our evaluation, we fixed $\alpha = 0.75$, $\beta = 0.75$.

We used more often our "IDF-based Query Expansion" however, based on the following procedure. First we formed the search term root set, composed of all terms included in the original query Q plus all indexing terms appearing in the *k* best ranked documents. The weight value for each term in this root set was computed as follows:

$$w'_j = \alpha \cdot I_Q(tj) \cdot idf_j + (\beta \cdot 1/k) \cdot \sum_{i=1}^{k} I_{D_i}(t_j) \cdot idf_j \quad (8)$$

$$\text{with } I_Q(t_j) = 1 \text{ if } t_j \in Q, \quad 0 \text{ otherwise}$$

where for term t_j, $idf_j = \ln(n/df_j)$ (the classical *idf* value) and $I_Q(t_j)$ (or $I_{Di}(t_j)$), an indicator function returning the value 1 if the term t_j belonging to the query Q (or the document D_i), otherwise the value was 0. In this weighting scheme, if a term appeared only in the original query Q, its weight would be $\alpha \cdot idf_j$, while a term appearing only in one document would have a weight of $(\beta/k) \cdot idf_j$.

The elements in the root set were then sorted in decreasing order according to their weight. To build the new query Q', we selected the top *m* search terms, and the weights attached to these selected terms in the new query were the same as those used in the root set. We thus used the same weighting scheme to select and weight the new search terms.

5 Data Fusion

In order to enhance the retrieval effectiveness and in an attempt to extract relevant documents with different features and characteristics, we decided to combine two or more result lists. During the indexing phase, we had already applied a combined approach when indexing either Chinese or Japanese corpus, using both the unigram and the bigram-based indexing scheme.

As a first data-fusion strategy, we considered the round-robin (RR) approach whereby one document in turn was selected from all individual lists and duplicates removed, retaining the highest ranking instances. Various other data fusion operators have been suggested [6], however the simple linear combination (denoted "Sum RSV") usually seemed to provide the best performance [6], or at least good overall performance [13], [14]. For a given set of result lists $i = 1, 2, \ldots, r$, this combined operator was defined as:

$$\text{Sum RSV}_i = \sum \text{RSV}_i \quad (9)$$

being the simple sum of all document scores (RSV_i) obtained by each search model.

	Mean average precision								
Stage 1	Chinese (50 queries)					Japanese (50 queries)			
Model	T	T	D	D	DN	T	T	D	D
Model 1	LM MTseg	Oka ubi	I(*n*)B2 ubi	I(*n*)B2 ubi	I(*n*)B2 ubi	IFB2 cha	IFB2 cha	IFB2 ubi	IFB2 ubi
#doc/#term	15 / 60 roc	15 / 100	10 / 150	**&** PB2 mts	10 / 150	5 / 70	5 / 70	15 / 140	**&** Oka cha
& IDFQE	0.2421	**0.2623**	**0.2839**	**0.2839**	**0.2873**	0.2856	0.2856	**0.2897**	
Model 2	Oka unibi	Oka mts	PB2 mts	Oka mts	PB2 mts	Oka ubi	IFB2 ubi	Oka cha	Oka unibi
#doc/#term	15 / 100	10 / 75	10 / 60	15 / 80	5 / 120	5 / 140	15 / 120	5 / 140	10 / 140
& IDFQE	0.2623	0.2596	0.2613	0.2562	0.2507	0.2844	**0.3003**	0.2530	**0.2940**
Model 3	I(*n*)B2 ubi			Oka ubi		Oka cha			PB2 cha
#doc/#term	15 / 100			10 / 150		5 / 130			5 / 90
& IDFQE	**0.2714**			0.2652		**0.2859**			0.2463
RR	0.2667	0.2718	0.2841	0.2781	0.2817	0.2922	0.3002	0.2773	0.2741
Sum RSV	0.2781	*0.2756*	**0.2904**	0.2883	**0.2904**	**0.3067**	**0.3108**	**0.2972**	**0.2985**
Norm RSV	0.2770	0.2744	0.2867	0.2862	0.2837	0.3030	0.3085	0.2901	0.2981
Z-score	***0.2783***	**0.2759**	*0.2889*	***0.2891****w*	*0.2794*	*0.2975w*	*0.3031w*	*0.2834*	*0.2894w*

Table 9. MAP with blind query expansion and various data fusion operators (Stage 1)
(Chinese and Japanese language, either with unigram & bigram (ubi) or Chasen/MT indexing strategy)

	Mean average precision					
Stage 1	Korean (50 queries)					Japanese
Model	T	T	D	D	DN	DN
Model 1	PB2 bigram	LM2a ham	PB2 bigram	PB2 bigram	PB2 ham	Okapi chasen
#doc/#term	5 / 90	5 / 100 roc	10 / 90	10 / 90	15 / 120	10 / 120
& IDFQE	**0.4128**	0.3437	**0.4403**	**0.4403**	0.4121	**0.3007**
Model 2	LM2a ham	LM2a bigram	LM bigram	LM bigram	LM bigram	
#doc/#term	5 / 100	5 / 70	5 / 40	5 / 40	5 / 70	
& Rocchio	0.3437	**0.4001**	0.4057	0.4057	**0.4172**	
Model 3				PB2 ham		Okapi unibi
#doc/#term				15 / 50		5 / 160
& IDFQE				0.4153		0.2691
Model 4				I(*n*)B2 bigram		
#doc/#term				15 / 140		
& IDFQE				0.4168		
RR	0.3825	0.3713	0.4305	0.4298	0.4218	0.2961
Norm RSV	**0.4155**	**0.4008**	**0.4382**	**0.4624**	*0.4431*	**0.2949**
Z-score	*0.4104*	*0.3920w*	*0.4362*	*0.4535w*	**0.4459**	*0.2938*

Table 10. MAP with blind query expansion and various data fusion operators (Stage 1) (Korean, either with bigram or HAM indexing strategy)

	Mean average precision – Chinese corpus								
Stage 2	NTCIR-3 (42 queries)			NTCIR-4 (59 queries)			NTCIR-5 (50 queries)		
Model	T	D	DN	T	D	DN	T	D	DN
Model 1	I(*n*)B2 ubi	PB2 mts	PB2 mts	I(*n*)B2 ubi	PB2 mts	PB2 mts	I(*n*)B2 ubi	PB2 mts	PB2 mts
#doc/#term	15 / 100	10 / 60	5 / 120	15 / 100	10 / 60	5 / 120	15 / 100	10 / 60	5 / 120
& IDFQE	**0.2695**	0.2994	0.3069	0.2296	**0.2482**	**0.2676**	**0.4257**	0.4007	**0.4656**
Model 2	Oka ubi	I(*n*)B2 ubi	I(*n*)B2 ubi	Oka ubi	I(*n*)B2 ubi	I(*n*)B2 ubi	Oka ubi	I(*n*)B2 ubi	I(*n*)B2 ubi
#doc/#term	10 / 100	10 / 150	10 / 150	10 / 100	10 / 150	10 / 150	10 / 100	10 / 150	10 / 150
& IDFQE	0.2662	**0.3133**	**0.3426**	0.2236	0.2481	0.2577	0.4125	**0.4103**	0.4366
Model 3	LM mts			LM mts			LM mts		
#doc/#term	15 / 60			15 / 60			15 / 60		
& Rocchio	0.2618			**0.2349**			0.3987		
RR	0.2653	0.3117	0.3341	0.2329	0.2530	0.2690	0.4209	0.4108	0.4612
Norm RSV	0.2803	0.3223	**0.3439**	**0.2370**	**0.2540**	0.2706	0.4358	0.4209	0.4704
Z-score	***0.2811***	***0.3247***	*0.3409*	*0.2358*	*0.2538*	***0.2714***	***0.4361***	***0.4240***	***0.4727***

Table 11. MAP with blind query expansion and various data fusion operators (Chinese, either with unigram & bigram (ubi) or MT segmentation indexing strategy)

	Mean average precision – Japanese corpus								
Stage 2	NTCIR-3 (42 queries)			NTCIR-4 (55 queries)			NTCIR-5 (47 queries)		
Model	T	D	DN	T	D	DN	T	D	DN
Model 1	IFB2 cha	Oka cha	Oka cha	IFB2 cha	Oka cha	Oka cha	IFB2 cha	Oka cha	Oka cha
#doc/#term	10 / 70	5 / 140	10 / 120	10 / 70	5 / 140	10 / 120	10 / 70	5 / 140	10 / 120
& IDFQE	**0.3699**	**0.3521**	**0.3956**	**0.3716**	0.3527	0.3512	**0.3968**	**0.3963**	**0.4347**
Model 2	IFB2 ubi	IFB2 ubi	Oka ubi	IFB2 ubi	IFB2 ubi	Oka ubi	IFB2 ubi	IFB2 ubi	Oka ubi
#doc/#term	15 / 120	15 / 140	5 / 160	15 / 120	15 / 140	5 / 160	15 / 120	15 / 140	5 / 160
& IDFQE	0.3274	0.3180	0.3915	0.3609	**0.3673**	**0.3563**	0.3673	0.3794	0.4326
RR	0.3541	0.3435	0.3971	0.3705	0.3624	0.3616	0.3911	0.3911	0.4443
Norm RSV	0.3652	0.3570	**0.4053**	0.3785	**0.3757**	0.3644	**0.4038**	0.4031	0.4476
Z-score	***0.3719***	***0.3601***	***0.4053***	***0.3795***	*0.3755*	***0.3650***	*0.4031*	***0.4023***	***0.4483***

Table 12. MAP with blind query expansion and various data fusion operators (Japanese, either with unigram & bigram (ubi) or Chasen segmentation indexing strategy)

	Mean average precision – Korean corpus								
Stage 2	NTCIR-3 (30 queries)			NTCIR-4 (57 queries)			NTCIR-5 (50 queries)		
Model	T	D	DN	T	D	DN	T	D	DN
Model 1	LM ham	LM big	LM big	LM ham	LM big	LM big	LM ham	LM big	LM big
#doc/#term	5 / 100	5 / 40	5 / 70	5 / 100	5 / 40	5 / 70	5 / 100	5 / 40	5 / 70
& Rocchio	0.2765	0.2721	0.3542	0.4114	0.4523	**0.4926**	**0.5012**	0.4913	0.5301
Model 2	PB2 big	PB2 big	PB2 ham	PB2 big	PB2 big	PB2 ham	PB2 big	PB2 big	PB2 ham
#doc/#term	5 / 90	10 / 90	15 / 120	5 / 90	10 / 90	15 / 120	5 / 90	10 / 90	15 / 120
& IDFQE	**0.3279**	0.2968	**0.3723**	**0.5081**	**0.5024**	0.4747	0.4988	**0.5055**	**0.5389**
Model 3		I(*n*)B2 big	Okapi big		I(*n*)B2 big	Okapi big		I(*n*)B2 big	Okapi big
#doc/#term		15 / 140	5 / 140		15 / 140	5 / 140		15 / 140	5 / 140
& IDFQE		0.2872	0.3637		0.4864	0.4868		0.5004	0.5270
Model 4		PB2 ham			PB2 ham			PB2 ham	
#doc/#term		15 / 50			15 / 50			15 / 50	
& IDFQE		**0.3129**			0.4545			0.4879	
RR	0.3043	0.2924	0.3789	0.4646	0.4772	0.4909	0.5105	0.5106	0.5433
Norm RSV	***0.3268***	***0.3242***	0.3779	***0.4938***	*0.5164*	0.5213	***0.5245***	*0.5248*	**0.5567**
Z-score	0.3223	0.3203	***0.3802***	0.4842	**0.5181**	***0.5224***	0.5244	**0.5249**	*0.5542*

Table 13. MAP with blind query expansion and various data fusion operators (Korean, either with bigram or HAM decompounding indexing strategy)

As a third data fusion strategy we normalized document scores within each collection through dividing them by the maximum score. As a variant of this normalized score merging scheme (denoted "Norm RSV"), we might normalize the document RSV_k scores within the *i*th result list, as follows:

$$\text{Norm } RSV_k = ((RSV_k - Min^i) / (Max^i - Min^i)) \quad (10)$$

As a fourth data fusion strategy, we would suggest merging the retrieved documents according to the Z-score, computed for each result list. Within this scheme, we would normalize retrieval status values for each document D_k provided by the *i*th result list, as computed by Equation 11.

$$\text{Z-score } RSV_k = \alpha_i \, . \, [((RSV_k - Mean^i) / Stdev^i) + \delta^i],$$
$$\delta^i = ((Mean^i - Min^i) / Stdev^i) \quad (11)$$

within which $Mean^i$ denotes the average of the RSV_k, $Stdev^i$ the standard deviation, and α_i (usually fixed at 1), used to reflect the retrieval performance of the underlying retrieval model.

In order to obtain a greater variety of relevant items, we used: 1) different IR models (Okapi, Language Model (LM) or DFR approaches), 2) various parameters settings when automatically expanding the original query, and 3) different data fusion operators when combining several runs.

As reported in Table 9, for Stage 1 when searching into the Chinese or Japanese corpora, we used three query formulations (T, D, DN) in varying combinations. Table 10 shows the same information for the Korean language in the last column, along with a run done on the Japanese collection. In these tables, official runs are indicated in italics. For Stage 2, Table 11 shows the main results obtained for the Chinese collections while in Tables 12 or 13, this same information is shown for the Japanese and Korean collections respectively.

Run name	MAP
UniNE-C-C-DN-01	0.2794
UniNE-C-C-T-02	0.2783
UniNE-C-C-T-04	0.2756
UniNE-C-C-D-03	0.2889
UniNE-C-C-D-05	0.2891
UniNE-J-J-DN-01	0.2938
UniNE-J-J-T-02	0.2975
UniNE-J-J-T-04	0.3031
UniNE-J-J-D-03	0.2894
UniNE-J-J-D-05	0.2834
UniNE-K-K-DN-01	0.4431
UniNE-K-K-T-03	0.4104
UniNE-K-K-T-05	0.3920
UniNE-K-K-D-02	0.4535
UniNE-K-K-D-04	0.4362

Table 14. MAP of our official runs (Stage 1)

Overall, the merging of two (or more) runs tended to result in improved MAP. This improvement was usually obtained however using either the norm RSV (Equation 10) or the Z-score (Equation 11) merging strategy, while the round-robin (RR) scheme tended to inhibit retrieval effectiveness. Applying a data fusion approach such as this resulted in very little enhancement however and seemed to be rather statistically insignificant. This finding is relatively similar across all three languages, although, for the Japanese language combining two or more runs may provide more consistent and beneficial results. On the other hand, a data fusion approach requires the manipulation of two (or more) inverted files and conducting two (or more) searches. To overcome this problem, we might consider merging two (or more) IR models during the indexing stage, as was done in the combined "unigram, & bigram" indexing scheme.

6 Official Results

Results from our official monolingual runs in Stage 1 are shown in Table 14. The evaluation of related runs can be found in Table 9 for Chinese and Japanese, or in Table 10 for the Korean corpus. The official runs are shown in italics in each of these tables.

7 Conclusion

Upon the completion of this evaluation campaign, we participated in building test collections for three Asian languages having a relatively large number of queries (namely 100), and analyzed various indexing and search strategies across four test collections (from NTCIR-3 to NTCIR-6).

Upon an analysis of the relative merits of various probabilistic and language models, what is clearly needed is a better understanding of the underlying mechanisms and reasoning behind good and bad performances on a query-by-query basis. The Okapi model for example usually results in the best MAP when searching Japanese corpora yet for the Korean collections the PB2 scheme clearly provides better results. Finally, the I(*n*)B2 model is usually the best for the Chinese corpora. For the moment however we are not able to provide a proper explanation of these facts.

From our evaluations (see Section 3), the language-independent bigram-based for the Korean language or the combined "unigram & bigram" indexing strategy for the Chinese and Japanese languages tend to offer better retrieval effectiveness than the more linguistically based indexing strategy (based on automatic segmentation for the Chinese (Mandarin Tools) or Japanese (ChaSen [11]) languages, or on a morphological analyzer (*Hangul Analyser Module*) for the Korean language).

Automatically expanding the submitted request by extracting terms from the *k* best-ranked documents (see Section 4) usually improves the MAP. This enhancement seems to be more effective for the DFR models. For example with the Japanese language, the Okapi is usually the best-performing search model before applying pseudo-relevance feedback. After query expansion, the I(*n*)B2 usually produces better MAP than the Okapi model. Moreover, including fewer search terms seems more effective in the case of a word-based indexing scheme than for a bigram-based or a combined "unigram & bigram" indexing scheme.

In order to extract more pertinent items from the various document collections, we have suggested applying a data fusion operator in order to combine two (or more) runs (see Section 5). Such a search strategy however requires that two (or more) searches need to be conducted, while also maintaining two (or more) inverted files. Given the relatively small improvements in MAP obtained, we are not however convinced that such a scheme would be really useful, at least from a commercial perspective.

Acknowledgments

The authors would like to thank the NTCIR-6 task organizers for their efforts in developing various test-collections. This research was supported in part by the Swiss National Science Foundation under Grants #200020-103420 and #200020-115866.

References

[1] Abdou, S., & Savoy, J. Report on CLIR task for the NTCIR-5 evaluation campaign. *Proceedings NTCIR-5*, NII, Tokyo, pp. 44-51, 2005.

[2] Abdou, S., & Savoy, J. Statistical and comparative evaluation of various indexing and search models. *Proceedings of AIRS-2006*, pp. 25-48, 2006.

[3] Amati, G., & van Rijsbergen, C.J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information System*, 20(4):357-389, 2002.

[4] Buckley, C., Singhal, A., Mitra, M., & Salton, G. New retrieval approaches using SMART. *Proceedings of TREC-4*, pp. 25-48, 1996.

[5] Chen, A., & Gey, F.C. Experiments on cross-language and patent retrieval at NTCIR-3 workshop. *Proceedings of NTCIR-3, 2003*.

[6] Fox, E.A., & Shaw, J.A. Combination of multiple searches. *Proceedings TREC-2*, pp. 243-249, 1994.

[7] Hiemstra, D. *Using language models for information retrieval*. CTIT Ph.D. Thesis, 2000.

[8] Hiemstra, D. Term-specific smoothing for the language modeling approach to information retrieval. *Proceedings ACM-SIGIR*, The ACM Press, pp. 35-41, 2002.

[9] Kishida, K., Chen, K.-H., Lee, S., Kuriyama, K., Kando, N., Chen, H.-H., & Myaeng, S.H. Overview of CLIR task at the sixth NTCIR workshop. *Proceedings of NTCIR-6*, Tokyo, 2007.

[10] Luk, R.W.P., & Kwok, K.L. A comparison of Chinese document indexing strategies and retrieval models. *ACM-Transactions on Asian Languages Information Processing*, 1(3):225-268, 2002.

[11] Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., & Asahara, M. *Japanese morphological analysis system ChaSen*. Technical Report NAIST-IS-TR99009, NAIST, 1999 (freely available at http://chasen.aist-nara.ac.jp/).

[12] Robertson, S.E., Walker, S., & Beaulieu, M. Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1):95-108, 2000.

[13] Savoy, J. Combining multiple strategies for effective monolingual and cross-lingual retrieval. *IR Journal*, 7(1-2):121-148, 2004.

[14] Savoy, J. Comparative study of monolingual and multilingual search models for use with Asian languages. *ACM-Transactions on Asian Languages Information Processing,* 4(3):163-189, 2005.