Report on CLIR Task for the NTCIR-4 Evaluation Campaign

Jacques Savoy
Institut interfacultaire d'informatique, Université de Neuchâtel
Pierre-à-Mazel 7, 2000 Neuchâtel, Switzerland
Jacques.Savoy@unine.ch

Abstract

This paper describes our first participation in an evaluation campaign involving three Asian languages (NTCIR-4). Our project has three objectives: 1) to compare the retrieval performances of eleven IR models used to carry out monolingual retrievals with these languages; 2) to analyze the relative merit of various freely available translation tools used to translate English-language topics into Chinese, Japanese or Korean; and 3) to evaluate the relative performance of the various merging strategies used to combine separate result lists extracted from a corpus written in English, Chinese, Japanese or Korean.

Keywords: CLIR, MLIR, data fusion, merging strategy.

1 Monolingual IR for Asian languages

In order to develop IR systems for Asian languages, many underlying assumptions previously made about European morphology must be revised, and different indexing strategies developed. This first section is organized as follows. Section 1.1 briefly describes the various corpora used in our evaluations (for more information, see [5]). Section 1.2 explains the main characteristics of the nine vector-space schemes and also the two probabilistic models used in our experiments. Section 1.3 provides an evaluation of various indexing and search strategies. Finally, Section 1.4 compares the relative merit of various data fusion operators.

1.1 Overview of NTCIR-4 test collection

Table 1 displays various statistics on corpora made available during the fourth NTCIR evaluation campaign (see also [5]). The Japanese collection is the largest, the English corpus is the second largest, with the Chinese and Korean corpus being the smallest. When comparing the number of distinct bigrams per article, the Chinese documents were usually quite large (363.4 bigrams/article), relative to the Korean (236.2) and Japanese (114.5) documents.

When analyzing the number of pertinent documents per topic, we only considered rigid assessments and thus in this paper only "highly relevant" and "relevant" items are seen as being relevant. A comparison of the number of relevant documents per topic, as shown in Table 1, indicates that for the Japanese collection the median number of relevant items per topic is 88, while for the Chinese corpus it is only 19. Clearly, the number of relevant articles is greater for the Japanese (7,137) and English (5,866) corpora, when compared to the Korean (3,131) or Chinese (1,318) collections. This fact may have an impact on some of our merging strategies (see Section 3).

Table 1 also provides an overview of the efficiency of the various search models, indicating the size of each collection in terms of storage space requirements and number of documents. The row labeled "# postings" indicates the number of indexing terms (word or bigram) in the inverted file, followed by the size of the inverted file and the time (user CPU time + system CPU time) needed to build the

	English	Chinese	Japanese	Korean
Size (in MB)	619 MB	490 MB	733 MB	370 MB
# of topics	58	59	55	57
Number rel. items	5,866	1,318	7,137	3,131
Median	35.5	19	88	43
# postings	524,788	2,704,517	804,801	320,431
Inverted file size	385 MB	1,187 MB	650 MB	530 MB
Building time	454.5 sec.	1,116.2 sec.	578.7 sec.	446.1 sec.
T mean query size	4.25 wd/query	5.8 bi/query	6.35 bi/query	5.58 bi/query
Search time per query	0.23 sec	0.183 sec	0.287 sec	0.187 sec
TDNC mean query size	34.25 wd/query	116.4 bi/query	28.7 bi/query	101.4 bi/query
Search time per query	0.433 sec	0.452 sec	0.492 sec	0.56 sec

Table 1. NTCIR-4 CLIR test collection statistics (rigid evaluation).

inverted file. To implement and evaluate these various search models, we used an Intel Pentium III/600 (memory: 1GB, swap: 2GB, disk: 6x35 GB). The average query size and time (in seconds) required to search for both short (T only) and long (TDNC) queries is shown in the lower rows (without blind query expansion).

1.2 Indexing and searching strategies

In our approach to these new test collections, we considered it important to evaluate the retrieval performance under various conditions, thus allowing us to draw some useful conclusions. In order to obtain this broader view, we decided to evaluate various indexing and search models. First we considered adopting a binary indexing scheme in which each document (or topic) was represented by a set of indexing terms (word for E, bigram for CJK), without any weight (IR model denoted "doc=bnn, query=bnn" or "bnn-bnn"). In order to weight the presence of each indexing term, we might account for the term occurrence frequency ("nnn-nnn") or we might also account for their frequency within the collection (or for idf). Moreover, when using cosine normalization, each indexing weight could vary within the range of 0 to 1 ("ntc-ntc").

Other variants might also be created. For example, the tf component could be computed as 0.5+0.5·[tf / max tf in a document] ("atn"). We might also consider that a term's presence in a shorter document provides stronger evidence than it would be in a longer document, leading to more complex IR models; i.e. the IR models denoted by "Lnu" [2] and "dtu" [10]. See the Appendix for the exact weighting formulation for the various IR models used in this paper.

In addition to previous models based on the vector-space approach, we also considered probabilistic models, such as the Okapi probabilistic model [7]. As a second probabilistic approach, we implemented the Prosit (or Deviation from Randomness) approach [1], based on the following indexing formula:

$$w_{ij} = (1 - Prob^2_{ij}) \cdot Inf^1_{ij}$$
$$\text{with } Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1)$$
$$\text{and } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean_l)/l_j)]$$
$$Inf^1_{ij} = -\log_2[1/(1+\lambda_j)] + tfn_{ij} \cdot \log_2[(\lambda_j+1)/\lambda_j]$$
$$\text{with } \lambda_j = tc_j / n$$

where w_{ij} represents the indexing weight attached to term t_j in document D_i, tc_j indicates the number of occurrences of term t_j in the collection and n the number of documents in the corpus.

For the English collection, we based the indexing process on the SMART stopword and stemmer. For the Asian languages, we indexed the documents using an overlapping bigram approach, an indexing scheme found effective for various Chinese collections [6], or during the last NTCIR campaign [3]. Based on this technique, the sequence "ABCD EFGH" would generate the following bigrams {"AB", "BC", "CD", "EF", "FG" and "GH"}. In our work, we generated these overlapping bigrams for Asian characters only, using spaces and other punctuation marks (collected for each language in their respective encoding) to stop bigram generation. Moreover, we did not split any words written with ASCII characters. Of course the most frequent bigrams may be removed before indexing. With the Chinese language for example, we defined and removed a list of 215 most frequent bigrams, for Japanese 105 bigrams and for Korean 80 bigrams. Finally for the Chinese language, we also evaluated the unigram (or character) indexing approach.

Before generating the bigrams for the Japanese documents, we removed all Hiragana characters, given that these characters are mainly used to write grammatical words (e.g., *doing*, *in*, *of*), and inflectional endings for verbs, adjectives and nouns. In our Japanese corpus, the Hiragana characters represented around 37.3% of the total, while 9.7% were Katakana, 46.3% were Kanji, and 6.7% ASCII (without counting half-width forms, punctuation or other drawing symbols).

1.3 Evaluation of various IR systems

To measure retrieval performance, we adopted a non-interpolated mean average precision (MAP), as computed by TREC_EVAL. To determine whether or not a given search strategy would be better than another, we based our statistical validation on the bootstrap approach [8]. Thus, in the tables appearing in this paper we have underlined statistically significant differences (two-sided non-parametric bootstrap test), based on those for which the mean difference had a significance level fixed at 5%.

We evaluated the various IR schemes under three topic formulations. First the queries were built using only the title (T), second using the descriptive (D) part and third using all topic logical sections (TDNC).

The mean average precision (MAP) as determined by the eleven search models is shown in Table 2 (for the English, Japanese and Korean collections), with the best performance for any given condition being shown in bold (these values were used as baseline for our statistical tests in Tables 2 and 3). Table 3 shows the performance achieved with the Chinese corpus using unigram (or characters) and bigram indexing schemes. Surprisingly, this data shows that the best retrieval scheme for short queries is not always the same as that for long topics. For Japanese and Chinese (bigram indexing) however, the best retrieval models are always the Okapi and the "Lnu-ltc" respectively. Based on our statistical testing, the difference in performance in not always significant (e.g., with the Japanese corpus, the differences between Okapi and "Lnu-ltc" models was only significant for T queries).

For the Chinese collection, when comparing character and bigram representations, it seemed that longer queries tended to perform better with bigram indexing. For T or D query constructions, the difference between character and bigram indexing

usually favored the bigrams approach (the performance of "Lnu-ltc" model with T queries is an exception). With the T queries and the Korean corpus, the binary indexing scheme ("bnn-bnn") has a surprisingly high retrieval performance when compared to the D or TDNC query formulations.

Moreover, we could also incorporate blind query expansion (or pseudo-relevance feedback) before presenting the result list to the user. In this study, we adopted Rocchio's approach [2] with α=4, β=0.75, γ=0.75, whereby the system was allowed to add *m* terms extracted from the *k* best-ranked documents from the original search. To evaluate this proposition, we used the Okapi and the Prosit probabilistic models. Table 4 summarizes the best results achieved for the English, Japanese and Korean language collections, while Table 5 shows the best retrieval performance for the Chinese collection (character or bigram indexing). In these tables, the rows labeled "Prosit" or "Okapi-npn" (baseline) indicate the mean average precision before applying this blind query expansion procedure. The rows starting with "#doc/#term" indicate the number of top-ranked documents and the number of terms used to enlarge the original query. Finally, the rows labeled "& Q exp." depict the mean average precision following blind query expansion (using the parameter setting specified in the previous row).

	Mean average precision								
	English (word, 58 queries)			Japanese (bigram, 55 queries)			Korean (bigram, 57 queries)		
Model	T	D	TDNC	T	D	TDNC	T	D	TDNC
Prosit	0.2977	0.2871	**0.3803**	0.2637	0.2573	0.3442	0.3882	0.3010	0.4630
Okapi-npn	**0.3132**	0.2992	0.3674	**0.2873**	**0.2821**	**0.3523**	0.4033	0.3475	**0.4987**
Lnu-ltc	0.3069	**0.3139**	0.3524	0.2701	0.2740	0.3448	**0.4193**	**0.4001**	0.4857
dtu-dtn	0.2945	0.2945	0.3126	0.2622	0.2640	0.3221	0.3830	0.3773	0.4397
atn-ntc	0.2808	0.2720	0.3417	0.2424	0.2405	0.3303	0.3604	0.3233	0.4202
ltn-ntc	0.2766	0.2908	0.3271	0.2735	0.2678	0.3265	0.3768	0.3494	0.4224
ntc-ntc	0.1975	0.2171	0.2559	0.2104	0.2087	0.2682	0.3245	0.3406	0.4133
ltc-ltc	0.1959	0.2106	0.2798	0.1868	0.1849	0.2596	0.3103	0.3205	0.4342
lnc-ltc	0.2295	0.2421	0.3235	0.1830	0.1835	0.2698	0.3231	0.3233	0.4616
bnn-bnn	0.1562	0.1262	0.0840	0.1743	0.1741	0.1501	0.1944	0.0725	0.0148
nnn-nnn	0.1084	0.1013	0.1178	0.1202	0.1099	0.1348	0.1853	0.1523	0.1711

Table 2. MAP for various IR models (E, J, and K monolingual).

	Mean average precision					
	Chinese (character, 59 queries)			Chinese (bigram, 59 queries)		
Model \ query type	T	D	TDNC	T	D	TDNC
Prosit	0.1452	0.0850	0.1486	0.1658	0.1467	0.2221
Okapi-npn	0.1667	0.1198	0.2179	0.1755	0.1576	0.2278
Lnu-ltc	**0.1834**	**0.1484**	0.2080	**0.1794**	**0.1609**	**0.2426**
dtu-dtn	0.1525	0.1103	0.1540	0.1527	0.1526	0.2239
atn-ntc	0.1334	0.0944	0.1699	0.1602	0.1461	0.2113
ltn-ntc	0.1191	0.0896	0.1371	0.1666	0.1556	0.2050
ntc-ntc	0.1186	0.1136	0.1741	0.1542	0.1507	0.1998
ltc-ltc	0.1002	0.0914	0.1905	0.1441	0.1430	0.2141
lnc-ltc	0.1396	0.1263	**0.2356**	0.1469	0.1438	0.2230
bnn-bnn	0.0431	0.0112	0.0022	0.0877	0.0781	0.0667
nnn-nnn	0.0251	0.0132	0.0069	0.0796	0.0687	0.0440

Table 3. MAP for various IR models (C monolingual).

	Mean average precision								
	English (word, 58 queries)			Japanese (bigram, 55 queries)			Korean (bigram, 57 queries)		
Model	T	D	TDNC	T	D	TDNC	T	D	TDNC
Prosit	0.2977	0.2871	0.3803	0.2637	0.2573	0.3442	0.3882	0.3010	0.4630
#doc/#term	10/125	10/175	5/140	10/300	10/100	10/125	5/120	3/130	10/175
& Q exp.	**0.3731**	**0.3513**	**0.3997**	**0.3396**	**0.3394**	**0.3724**	0.4875	0.4257	0.5126
Okapi-npn	0.3132	0.2992	0.3674	0.2873	0.2821	0.3523	0.4033	0.3475	0.4987
#doc/#term	10/120	10/10	10/120	10/15	5/100	5/175	10/160	5/140	10/150
& Q exp.	0.3594	0.3181	0.3727	0.3259	0.3331	0.3640	**0.4960**	**0.4441**	**0.5154**

Table 4. MAP with blind query expansion (E, J, and K monolingual).

	Mean average precision					
	Chinese (character, 59 queries)			Chinese (bigram, 59 queries)		
Model	T	D	TDNC	T	D	TDNC
Prosit	0.1452	0.0850	0.1486	0.1658	0.1467	0.2221
#doc/#term	10/0/125	10/0/75	3/0/10	10/0/175	10/0/100	5/0/20
& Qexp.	0.1659	0.1132	0.1624	**0.2140**	**0.1987**	**0.2507**
Okapi-npn	0.1667	0.1198	0.2179	0.1755	0.1576	0.2278
#doc/#term	10/0/10	10/0/10	10/0/60	5/0/25	5/0/100	5/0/60
& Qexp.	**0.1884**	**0.1407**	**0.2213**	0.2004	0.1805	0.2331

Table 5. MAP with blind query expansion (C monolingual).

	Mean average precision								
	Chinese (bigram/unigram, 59 q.)			Japanese (bigram, 55 queries)			Korean (bigram, 57 queries)		
Model	T	D	TDNC	T	D	TDNC	T	D	TDNC
#doc/#term	5/0/30	10/0/100	10/0/60	10/0/200	10/0/75	10/0/350	10/0/100		3/0/30
Prosit	*0.2007*	0.1987	***0.2450***	0.3388	*0.3390*	0.3688	0.4868		0.4657
#doc/#term	5/0/100	10/0/100		5/0/10	10/0/150	10/0/150	3/0/30	5/0/20	10/0/40
Okapi-npn	0.1987	0.1758		0.3181	0.3324	0.3624	*0.4654*	0.4335	**0.5141**
#doc/#term	3/0/75	5/0/125		10/0/350	5/0/75	10/0/200	10/0/300		
Lnu-ltc	0.1824	0.1711		0.2879	0.2884	0.3545	0.4500		
#doc/#term	10/0/40	5/0/60		10/0/350			5/0/5	5/0/10	
ltn-ntc	0.1780	0.1898		0.2786			0.4303	0.3946	
#doc/#term	10/0/10	3/0/10							
Okapi-npn	0.1884	0.1394	<- unigram search model						
#doc/#term	3/0/75	3/0/60							
Lnu-ltc	0.1926	0.1592	<- unigram search model						
Round-rob.	0.1903	0.1778		0.3283	0.3385	0.3679	0.4737	0.4260	*0.5047*
SumRSV	0.2103	0.1947		0.3455	0.3420	0.3739	0.5044	0.4391	0.5030
NormRSV	0.2120			*0.3486*	0.3444	0.3746	**0.5084**	*0.4431*	0.5045
Z-score	***0.2135***	0.1996		0.3498	0.3458	**0.3755**	0.5074	0.4442	0.5023
Z-score W	0.2120	***0.2011***		**0.3513**	***0.3484***	*0.3728*	*0.5078*	**0.4471**	0.5058

Table 6. MAP with various data fusion schemes (official runs in italics).

From the data shown in Tables 4 and 5, we could infer that the blind query expansion technique improves the mean average precision, and this improvement is usually statistically significant (value underlined in the table). When comparing both probabilistic models, this strategy seems to perform better with the Prosit than with the Okapi model. In addition, the percentage enhancement is greater for short topics than for longer ones. For example, in the Japanese collection with the Prosit model and T topics, blind query expansion improved mean performance, ranging from 0.2637 to 0.3396 (+28.8% in relative effectiveness), as compared to 0.3442 to 0.3724 (+8.5%) for TDNC topics.

1.4 Data fusion

As an additional strategy to enhance retrieval effectiveness, we considered adopting a data fusion approach that combined two or more result lists provided by different search models. In this case, we viewed each IR model as a distinct and independent source of evidence of document relevance. As a first data fusion strategy, we considered the round-robin ("RR") approach whereby we took one document in turn from all individual lists and removed duplicates, keeping the most highly ranked instance. Various other data fusion operators have been suggested [4], however the simple linear combination (denoted "SumRSV") usually seemed to provide the best performance [9], [4]. Given a set of results lists $i = 1, 2, \ldots, r$, this combined operator is defined as $\text{SumRSV} = \text{SUM}(\alpha_i \cdot \text{RSV}_i)$, in which the value of α_i (fixed at 1 for all result lists in our experiments) may be used to reflect retrieval performance differences between IR models.

Unfortunately document scores cannot usually be directly compared, thus as a third data fusion strategy we normalized document scores within each collection through dividing them by the maximum score, denoted "NormMax" (i.e. the document score of the retrieved record in the first position). As a variant of this normalized score merging scheme (denoted "NormRSV"), we might normalize the document RSV_k scores within the *i*th result list, according to Equation 1.

$$\text{NormRSV}_k = ((\text{RSV}_k - \text{Min}^i) / (\text{Max}^i - \text{Min}^i)) \quad (1)$$

As a fourth and new data fusion strategy, we suggest merging the retrieved documents according to the Z-score, computed for each result list. Within this scheme, for the *i*th result list, we needed to compute the average of the RSV_k (denoted Mean^i) and the standard deviation (denoted Stdev^i). Based on these values, we would then normalize the retrieval status

value for each document D_k provided by the *i*th result list, by computing the following formula:

$$\text{Z-score } RSV_k = \sum_i \alpha_i \cdot [((RSV_k - Mean^i) / Stdev^i) + \delta^i],$$
$$\delta^i = ((Mean^i - Min^i) / Stdev^i) \quad (2)$$

within which the value of δ^i is used to generate only positive values, and α_i (usually fixed at 1) is used to reflect the retrieval performance of the underlying retrieval model. When the coefficients α_i are not all fixed at 1, the data fusion operator is denoted as "Z-score W".

Table 6 shows the mean average precision (MAP) obtained from the Chinese, Japanese and Korean collections, for each of the T, D and TDNC queries. In this table, the round-robin ("RR") scheme was to serve as baseline for our statistical testing. From this data, we could see that combining two or more IR models might sometimes improve retrieval effectiveness. Moreover, a linear combination ("SumRSV") usually resulted in good performance, and the Z-score scheme tended to produce the best performance. In Table 6, under the heading "Z-score W", we attached a weight of 2 to the Prosit model, 1.5 to the Okapi and 1 to other IR models. However, combining separate result lists did not always enhance the performance, as shown by the Korean collection with TDNC queries. It is difficult however to predict which data fusion operator would produce the best result, even when a particular data fusion scheme improved performance over single runs. Our experiments also indicate that combining short queries results in more improvement than do longer topics.

Results from some of our official monolingual runs are shown in Table 6 and are indicated in italics. For the Chinese monolingual task, the UniNE-C-C-T-05 and UniNE-C-C-T-05 are shown in the second column, UniNE-C-C-D-03 in the third column, and UniNE-C-C-TDNC-02 in the forth. For the Japanese monolingual task, the UniNE-J-J-T-04 run is shown in the fifth column, the UniNE-J-J-D-05 and UniNE-J-J-D-02 runs in the sixth column, and the UniNE-J-J-TDNC-01 run in the seventh column. For the Korean language, the UniNE-K-K-T-04 and UniNE-K-K-D-03 runs are shown in the eighth column, the UniNE-K-K-D-05 run in the ninth column and the UniNE-K-K-TDNC-01 run in the last column.

2 Bilingual IR

In order to retrieve information written in one Far-East language for a topic written in English, we based our approach on freely available resources that automatically provide translations in Chinese, Japanese or Korean languages. In this study, we chose four different machine translation (MT) systems and two machine-readable bilingual dictionaries (MRDs) to translate the topics:

	Mean average precision								
	Chinese (bigram, 59 queries)			Japanese (bigram, 55 queries)			Korean (bigram, 57 queries)		
Model	T	D	TDNC	T	D	TDNC	T	D	TDNC
Okapi-npn	0.1755	0.1576	0.2278	0.2873	0.2821	0.3523	0.4033	0.3475	0.4987
Babylon 1	0.0458	0.0459	0.0643	0.0946	0.1255	0.1858	0.1015	0.0628	0.0706
Babylon 2	0.0441	0.0434	0.0607	0.0899	0.1202	0.1766	0.0948	0.0625	0.0660
Babylon 3	0.0473	0.0412	0.0651	0.0911	0.1172	0.1651	0.0925	0.0611	0.0627
EvDict	0.0465	0.0532	0.0753	n/a	n/a	n/a	n/a	n/a	n/a
WorldLing	0.0794	0.0702	0.1109	0.1951	**0.1972**	0.2385	0.1847	0.1745	0.2694
Babelfish	0.0360	0.0337	0.0507	0.1952	**0.1972**	0.2390	**0.1855**	**0.1768**	**0.2739**
InterTrans	n/a	n/a	n/a	0.0906	0.0888	0.1396	n/a	n/a	n/a
FreeTrans	0.0665	0.0643	0.0967	n/a	n/a	n/a	n/a	n/a	n/a
Combined	Lingo+EvDict			Lingo+Babylon 1			Lingo+Babelfish		
with Okapi	**0.0854**	**0.0813**	**0.1213**	**0.2174**	0.1951	**0.2550**	0.1848	**0.1768**	0.2706
with Prosit	0.0817	0.0728	0.1133	0.1973	0.1897	0.2508	0.1721	0.1475	0.2409

Table 7. MAP for various query translation approaches (Okapi model).

	Mean average precision								
	Chinese (bigram, 59 queries)			Japanese (bigram, 55 queries)			Korean (bigram, 57 queries)		
Model	T	D	TDNC	T	D	TDNC	T	D	TDNC
Okapi-npn	0.0854	0.0813	0.1213	0.2174	0.1951	0.2550	0.1848	0.1768	0.2706
#doc/#term & Q exp.	5/10/G 0.1039	5/10/G 0.1003	5/10/H 0.1290	10/10/H **0.2733**	5/10/H 0.2185	5/10/H 0.2669	5/10/H **0.2397**	10/10/M **0.2139**	5/10/G 0.2882
Prosit	0.0817	0.0728	0.1133	0.1973	0.1897	0.2508	0.1721	0.1475	0.2409
#doc/#term & Q exp.	5/10/E **0.1213**	10/10/I **0.1057**	5/10/G **0.1644**	10M 0.2556	10I **0.2600**	10M **0.3065**	10/10/I 0.2326	10/10/I 0.2098	10/10/I **0.2968**

Table 8. MAP for blind query expansion on translated queries (Okapi or Prosit).

BABELFISH	babel.altavista.com/translate.dyn
FREETRANSLATION	www.freetranslation.com
INTERTRAN	www.tranexp.com:2000/InterTran
WORLDLINGO	www.worldlingo.com
EVDICT	www.samlight.com/ev/
BABYLON	www.babylon.com

For the Babylon bilingual dictionary, we submitted search keywords word-by-word. In response to each word submitted, the Babylon system provided not only one but several translation terms (in an unspecified order). In our experiments, we decided to pick the first available translation (labeled "Babylon□1"), the first two (labeled "Babylon□2") or the first three (labeled "Babylon□3").

Table□7 shows mean average precision when translating English topics employing our two MRDs, the four MT systems and the Okapi model. This table also contains the retrieval performance for manually translated topics, with the first row ("Okapi-npn") being used as a baseline. Since some translation devices were not able to provide a translation for each language, Table□7 indicates these missing entries as "n/a". Compared to our previous work with European languages [9], machine translated topics provided generally poor performance levels when compared to manually translated topics. Based on the T queries and the best single query translation resource, we only obtained 45.2% of the performance level achieved by a monolingual search for the Chinese language (0.0794 vs. 0.1755), 67.9% for the Japanese (0.1952 vs. 0.2873) or 46% for the Korean language (0.1855 vs. 0.4033). Moreover, the differences in mean average precision were always statistically significant and favored manual topic translation approaches.

The Babelfish MT system seemed to produce the best translated topics for the Japanese and Korean languages, and WorldLingo for the Chinese. The poor performance displayed by Babelfish when translating the Chinese language seemed to be caused by a conversion problem (the Babelfish output format is in simplified Chinese, and we needed the topic in BIG5 encoding).

To improve the retrieval performance of translated topics, we developed three possible strategies. First, we combined the translation provided by two translation tools. For the Japanese language, we concatenated the results supplied by WorldLingo with those of "Babylon□1", and for Korean, we combined the translations provided by WorldLingo with those of Babelfish. As shown in the last two rows of Table□7, this combined translation strategy seemed to enhance retrieval effectiveness for the Chinese and Japanese languages, but not for Korean.

Our second attempt to improve performance was to apply a blind query expansion to the combined translated topics. As shown in Table□8, this technique clearly enhanced retrieval effectiveness when the Okapi or the Prosit probabilistic models were used. As for monolingual IR (see Table□4), the results achieved by the Prosit system after pseudo-relevance feedback were usually better than those obtained by the Okapi search model. Surprisingly, for T queries in the Japanese corpus, the Okapi with blind query expansion achieved a performance level of 0.2733 (or 95.1% of the monolingual performance, however without blind query expansion). When compared to other bilingual runs, our approach seemed very attractive, at least for the Chinese and Japanese languages.

As a third strategy for enhancing retrieval effectiveness, we considered adopting a data fusion approach that combined two or more result lists provided by different search models (as was done in the monolingual search, see Section 1.4).

3 Multilingual IR

In this section, we will investigate the situation where users write a topic in English in order to retrieve relevant documents in English, Chinese, Japanese and Korean (CJE and CJKE context). To deal with this multi-language barrier we based our approach on bilingual IR systems, as described in the previous section. Thus, the different collections were indexed separately and, once the original or a translated request was received, a ranked list of retrieved items was returned. From these lists we needed to produce a unique ranked result list, using a merging strategy described further on in this section.

As a first approach, we considered the round-robin ("RR") method, whereby we took one document in turn from all individual lists. As a second merging approach, we took the document score into account, denoted as RSV_k for document D_k. This strategy, called raw-score merging, produces a final list sorted by document score, as computed by each collection. As a third scheme, we could normalize the RSV_k using the document score of the retrieved record in the first position ("MaxRSV") or by using Equation□1 ("NormRSV").

As a fifth merging scheme, we suggested a biased round-robin approach which extracts not just one document per collection per round, but one document from both the English and Chinese collections and two from the Japanese and Korean. Such a merging strategy exploits the fact that the Japanese and Korean corpora possess more articles than do the English or the Chinese collections (see Table□1). Finally, we could use our new Z-score (see Section□1.4 and Equation□2) to define a comparable document score across collections. Under the label "Z-score W", we assigned a weight of 2 for the Japanese and Korean result lists and 1 for the English and Chinese runs.

Table□9 shows the retrieval effectiveness of the various merging strategies. The top part of this table shows the mean average precision obtained independently for each language (based on a smaller number of queries) and using the Prosit, Okapi, and "Lnu-ltc" search models along with query expansion or a data fusion approach for the various bilingual searches (based on the Z-score scheme and denoted

"DF-Zscore(k)", with *k* indicating the number of merged runs). In the last three columns of Table 9, we evaluated multilingual runs using manually translated topics in order to estimate decreases in retrieval effectiveness due to the automatic query translation strategies. In this table, the round-robin merging strategy served as a baseline upon which the statistical tests were based.

The data depicted in Table 9 also indicates that only a few runs produced retrieval effectiveness that could be viewed as statistically superior to that of the round-robin baseline. As a first approach, both simple, normalized merging schemes ("MaxRSV" or "NormRSV") provided reasonable performance levels, with the "NormRSV" merging scheme having a slight advantage. In our case, the raw-score approach did not result in interesting retrieval effectiveness and performance decreases were usually statistically significant when compared to the round-robin scheme. In this experiment we merged result lists obtained by various IR models, however the resulting document scores were incomparable, thus rendering the raw-score approach ineffective. Also, our biased round-robin scheme did not perform better when compared to the simple round-robin version (moreover, it is difficult a priori to know whether a given corpus will really contain more relevant items than another). Both the Z-score and the weighted Z-score (with [illegible] for the English and Chinese corpora and 2 for both the Japanese and Korean languages) usually provided better performance levels than the round-robin approach (the difference in performance was not however always statistically significant).

The difference in performance between manually and automatically translated queries was relatively important. For CJE multilingual retrieval and T queries, the best automatic run achieved a mean average precision of 0.1719 compared to 0.2370 (or 27.5% of difference in relative performance). When compared with CJKE multilingual search and T queries, the difference was larger (0.1446 vs. 0.2549, or 43.3%).

The top section of Table 9 shows three of our official monolingual English runs, indicated in italics (runs UniNE-E-E-T-03, UniNE-E-E-D-04, and UniNE-E-E-TDNC-01). Four of our official runs for the CJE multilingual task are also listed (the UniNE-E-CJE-T-04 and UniNE-E-CJE-T-05 runs midway down the second column, and the UniNE-E-CJE-D-02 and UniNE-E-CJE-D-03 runs midway down the third column). For the CJKE multilingual task, Table 9 shows two of our official runs, the UniNE-E-CJKE-T-04 and UniNE-E-CJKE-T-05 runs in the bottom of column two, (in our official UniNE-E-CJKE-D-02 and UniNE-E-CJKE-D-03 runs, Korean corpus searches was based on DNC queries, thus they performed better than those depicted in Table 9).

	Mean average precision					
	Queries automatically translated			Queries manually translated		
	T	D	TDNC	T	D	TDNC
English (on 58 queries)	Prosit[illegible]0/30	Prosit[illegible]0/15	Prosit[illegible]/50	Prosit[illegible]0/125	Prosit[illegible]0/75	Prosit[illegible]/40
	0.3576	*0.3169*	*0.3856*	0.3731	0.3513	0.3997
Chinese (on 59 queries)	DF-Zscore(2)	Prosit[illegible]0/125	Prosit[illegible]/30	Prosit[illegible]0/175	Prosit[illegible]0/100	Lnu[illegible]/125
	0.1000	0.1057	0.1596	0.2140	0.1987	0.2516
Japanese (on 55 queries)	DF-Zscore(4)	DF-Zscore(2)	DF-Zscore(2)	Prosit[illegible]0/300	Prosit[illegible]0/100	Prosit[illegible]0/125
	0.2752	0.2628	0.2896	0.3396	0.3394	0.3724
Korean (on 57 queries)	DF-Zscore(2)	DF-Zscore(2)	DF-Zscore(2)	Okapi[illegible]0/60	Okapi[illegible]/40	Okapi[illegible]0/50
	0.2410	0.2075	0.2926	0.4960	0.4441	0.5154
Merging strategy CJE						
Round-robin (baseline)	0.1564	0.1484	0.1913	0.2204	0.2114	0.2500
Raw-score	0.1307	0.0521	0.1102	0.2035	0.1981	0.2100
MaxRSV	0.1654	0.1473	0.1936	0.2222	0.2180	0.2415
NormRSV (Eq.[illegible])	*0.1685*	*0.1604*	0.2006	0.2281	0.2195	0.2541
Biased RR E1/C2/J2	*0.1413*	0.1343	0.1736	0.2290	0.2198	0.2569
Z-score (Eq. 2)	0.1624	*0.1575*	**0.2028**	0.2293	0.2243	0.2596
Z-score W E1/C1/J2	**0.1719**	**0.1645**	0.1978	**0.2370**	**0.2293**	**0.2625**
Merging strategy CJKE						
Round-robin (baseline)	0.1419	0.1322	0.1800	0.2371	0.2223	0.2608
Raw-score	0.1033	0.0382	0.0861	0.1564	0.1513	0.1657
MaxRSV	0.1411	0.1285	0.1816	0.2269	0.2192	0.2506
NormRSV (Eq.[illegible])	*0.1437*	0.1392	0.1799	0.2481	0.2278	0.2706
Biased RR E1/C1/J2/K2	*0.1320*	0.1220	0.1672	0.2431	0.2266	0.2645
Z-score (Eq. 2)	**0.1446**	**0.1398**	**0.1880**	0.2483	0.2360	0.2716
Z-score W E1/C1/J2/K2	0.1332	0.1377	0.1763	**0.2549**	**0.2380**	**0.2735**

Table 9. MAP of various merging strategies for CJE collection (medium) and CJKE collection (bottom), (official runs in italics).

Conclusion

Based on our evaluations, we have shown that when indexing Asian languages based on bigrams, the IR models providing the best retrieval performance levels are the "Lnu-ltc" vector-space model or the Okapi probabilistic model (see Tables 2 or 3). To improve retrieval effectiveness, a blind query expansion is a worthwhile approach, especially when processing short queries and using the Prosit IR model (see Tables 4 or 5). In order to further improve retrieval effectiveness a data fusion approach could also be considered, although this technique would require additional computational resources (Table 6).

When analyzing the performance of bilingual searches, our results were contrary to those found for certain European languages [9], with the number and quality of freely available translation resources being questionable. When translating the topics from English into Chinese, Japanese or Korean language, the overall retrieval effectiveness decreases more than 30% for the Japanese, and more than 50% for the Chinese and Korean languages (see Table 7). To improve this poor performance, we might concatenate two (or more) translations (see the last two rows of Table 7), employ a blind query expansion approach (see Table 8), and a data fusion approach.

When evaluating various merging strategies (see Table 9) using different query sizes, it appears that the Z-score merging procedure tends to produce interesting retrieval effectiveness when merging ranked lists of retrieved items provided by separate collections.

Acknowledgments

This research was supported by the Swiss NSF (Grants #21-66742.01 and #20-103420/1).

References

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

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

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

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

[5] K. Kishida, K.-H. Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, S.H. Myaeng, K. Eguchi. Overview of CLIR Task at the Forth NTCIR Workshop. *Proceedings of NTCIR-4*, 2004.

[6] R.W.P. Luk, K.L. Kwok. A comparison of Chinese document indexing strategies and retrieval models. *ACM-TOALIP*, 1(3):225-268, 2002.

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

[8] J. Savoy. Statistical inference in retrieval effectiveness evaluation. *IP&M*, 33(4):495-512, 1997.

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

[10] A. Singhal, J. Choi, D. Hindle, D.D. Lewis, F. Pereira. AT&T at TREC-7. *Proceedings of TREC-7*, pp. 239-251, 1999.

Appendix

In Table A.1, w_{ij} represents the indexing weight assigned to term t_j in document D_i. To achieve this, n indicates the number of documents and nt_i the number of distinct indexing units (bigrams or terms) included in D_i representation. In our experiments, we assigned values to the constant b as follows: 0.5 for both the Chinese and Japanese collections, 0.55 for the English, and 0.75 for the Korean, while we fixed the constant k_1 at 1.2, *avdl* at 500, *pivot* at 100, and the slope at 0.1. For the Prosit model, $c = 0$ for the Japanese and Korean corpus, $c = 0.1$ for the English, and $c = 0.5$ for the Chinese. These values were chosen because they usually result in a better retrieval effectiveness. The value *mean dl* was fixed at 151 for the English, 480 for the Chinese, 144 for the Japanese, and 295 for the Korean corpus.

bnn	$w_{ij} = 1$	npn	$w_{ij} = tf_{ij} \cdot \ln[(n-df_j)/df_j]$	ltn	$w_{ij} = [\ln(tf_{ij}) + 1] \cdot idf_j$
nnn	$w_{ij} = tf_{ij}$	lnc	$w_{ij} = \dfrac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} (\ln(tf_{ik}) + 1)^2}}$	ntc	$w_{ij} = \dfrac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}}$
atn	$w_{ij} = idf_j \cdot [0.5 + 0.5 \cdot tf_{ij} / \max tf_i]$			dtn	$w_{ij} = [\ln[\ln(tf_{ij}) + 1] + 1] \cdot idf_j$
Lnu	$w_{ij} = \dfrac{\left(1 + \ln(tf_{ij})\right) / \left(\ln(\text{mean tf}) + 1\right)}{(1 - \text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$			ltc	$w_{ij} = \dfrac{(\ln(tf_{ij}) + 1) \cdot idf_j}{\sqrt{\sum_{k=1}^{t} ((\ln(tf_{ik}) + 1) \cdot idf_k)^2}}$
Okapi	$w_{ij} = \dfrac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}}$			dtu	$w_{ij} = \dfrac{\left(\ln\left(\ln(tf_{ij}) + 1\right) + 1\right) \cdot idf_j}{(1 - \text{slope}) \cdot \text{pivot} + (\text{slope} \cdot nt_i)}$

Table A.1. Weighting schemes.