Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages

Jacques Savoy[1]
Université de Neuchâtel, Switzerland

Abstract

Based on the NTCIR-4 test-collection, our first objective is to present an overview of the retrieval effectiveness of nine vector-space and two probabilistic models when performing monolingual searches in the Chinese, Japanese, Korean and English languages. Our second goal is to analyze the relative merits of using various automated and freely available tools to translate English-language topics into Chinese, Japanese or Korean, and then submit the resultant query to retrieve pertinent documents written in one of these three Asian languages. We also demonstrate how bilingual searches could be improved by applying both combined query translation strategies and data fusion approaches. Finally, we address basic problems related to multilingual searches in which queries written in English are used to search documents written in the English, Chinese, Japanese and Korean languages.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval.

General Terms: Experimentation, Measurement, Algorithms.

Additional Keywords and Phrases: Multilingual Information Retrieval, Cross-Language Information Retrieval, Natural Language Processing with Asian Languages, Results Merging, Chinese Language, Japanese Language, Korean Language, Search Engines with Asian Languages.

1 Monolingual IR for Asian languages

During the last few years, there has been increasing interest in Asian languages, particularly Chinese (C), Japanese (J) and Korean (K). Given the growing number of Internet pages and sites available in these languages, along with an ever expanding number of online users[2] working with them, a better understanding of the automated procedures used to process them is clearly needed. These Asian languages also represent various external differences that, compared to European languages, present the IR community very interesting challenges.

Firstly, while the Latin alphabet consists of only 26 characters (or 33 in the Cyrillic and 28 in the Arabic alphabets), standard Asian languages require quite a larger number of characters (around 13,000 for the Chinese BIG5 encoding system, around 8,200 for the Korean and 8,800 for the Japanese). Secondly, when processing Far-East languages, the implicit assumption that one byte corresponds to one character is no longer valid. These facts thus lead to additional challenges for anyone using typical Unix functions such as `wc`, `sort` and `grep`, and generally entail the use of more complex input and output methods [Lunde 1998].

[1] Author address: Groupe d'informatique, Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel, Switzerland. E-mail: Jacques.Savoy@unine.ch

[2] See the Web site at http://global-reach.biz/globstats/

Thirdly, the typical Chinese sentence consists of a continuous string of characters (or more precisely ideographs[3]) without any delimiting spaces separating them. As such, finding words within such a continuous string becomes a major problem, one that has to be resolved before tackling various other problems such as linguistic analysis, machine translation or information retrieval. Fourthly, the Chinese language may be written using one of two main character formats. These are the traditional (usually encoded using the BIG5 standard) and the simplified (using the GB standard system [Lunde 1998]), not to mention considerable orthographic variations encountered when spelling foreign names [Halpern 2002][4]. Fifthly, in Japanese there are four writing systems used, namely the Hiragana syllabic character set (representing around 37.3% of the total number of characters[5]); the Katakana (9.7%, a syllabic character set used mainly to write words of foreign origin such as "computer"; foreign names like "MacIntosh" or onomatopoeic words like "buzz"); the Kanji (corresponding to Chinese characters and making up 46.3% of characters used); and finally ASCII characters (about 6.7%, used to write numbers or company names such as "Honda").

Added to these visual differences between the European and Asian languages are morphological differences [Sproat 1992]. On the basis of morphological information or word structure, the languages studied in our evaluations can be broadly grouped into three different types, based on Bloomfield's classification [1933][6]. Firstly, English, Latin and most other European languages are inflectional languages within which certain distinct features are used to create single or fairly unified suffix formats added to a given stem (inflectional suffixes such as "-s" in "runs" or derivational suffixes like "-ment" in "establishment"). Secondly, Chinese belongs to an isolated language family in which the vast majority of words are invariable, meaning that in IR system stemming procedures would play a less important role. Lastly, while both the Japanese or Korean languages may be considered as members of the agglutinative languages family in which various affixes are added to a given stem, they may also belong to a separate class that has neither a clear nor close relationship with any other language.

Given these visual and morphological differences between Indo-European and Asian languages, it is important to verify whether or not the efficient search models already developed for European languages will perform as well with Asian languages. The first section in this paper addresses this question and is organized as follows. Section 1.1 briefly describes the various corpora used in our evaluations. Section 1.2 explains the main characteristics of the nine vector-space schemes and the two probabilistic IR models. Section 1.3 presents the indexing strategies used in our experiments. Section 1.4 provides an evaluation of various indexing and search strategies while Section 1.5 evaluates a pseudo-relevance feedback approach intended to improve retrieval effectiveness. Finally, Section 1.6 compares the relative merit of various data fusion operators.

---

[3] Also referred to as pictographs or logographs, depending on their etymology.

[4] Each natural language has some of these orthographic variations (such as "color" and "colour" in English). However, the main differences are related to homophones involving proper names. For example, Stephenson, the inventor of the steam engine, and Stevenson, the author, have the same pronunciation, and both names may be written identically in Japanese, Chinese of Korean languages.

[5] Without counting half-width forms, punctuation or other graphic or drawing symbols.

[6] We may also classify languages according to their word order (namely the order in which Subject, Verb, and Object appear in a normal sentence), being SVO for the English and Chinese languages, and SOV for the Japanese and Korean languages. Word order does not however usually play an important role in various IR systems, at least not in those used in this paper.

1.1 Overview of NTCIR-4 test-collection

The corpora used in our experiments were put together during the fourth NTCIR[7] evaluation campaign [Kishida et al. 2004a]. Created to promote information retrieval studies on Asian languages, this test-collection includes various newspapers written in four different languages. The English collection is taken from the *Mainichi Daily News* (Japan), *Taiwan News*, *China Times English News* (Taiwan), the *Xinhua News Service* (China), the *Korean Times* and the *Hong Kong Standard*. The Chinese collection contains news extracted from the *United Daily News*, *China Times*, *China Times Express*, *Commercial Times*, *China Daily News* and *Central and Daily News*. These documents were written in Mandarin using the traditional Chinese character set. The Japanese collection contains articles taken from the *Mainichi* and *Yomiuri* newspapers (Japan), while the Korean corpus was extracted from both the *Hankookilbo* and *Chosunilbo* newspapers (Korea).

	English	Chinese	Japanese	Korean
Size (in MB)	619 MB	490 MB	733 MB	370 MB
# of documents	347,376	381,375	593,636	254,438
Publication year	1998-1999	1998-1999	1998-1999	1998-1999
Encoding	ASCII	BIG5	EUC-JP	EUC-KR
Number of distinct indexing words or bigrams / document				
Mean	96.6	363.4	114.5	236.2
Standard deviation	61.9	219.9	97.0	146.2
Median	82	326	90	209
Maximum	2,052	5,935	5,232	3,762
Minimum	1	1	1	2
Number of topics	58	59	55	57
Number of relevant items	5,866	1,318	7,137	3,131
Mean relevant items / topic	101.138	22.339	129.764	54.9298
Standard deviation	130.785	13.502	119.56	40.8507
Median	35.5	19	88	43
Maximum	642 (Q#47)	61 (Q#18)	548 (Q#57)	171 (Q#9)
Minimum	5 (Q#2)	3 (Q#9)	6 (Q#4)	3 (Q#52)

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

Table 1 compares the various sizes of these corpora, ranking the Japanese collection as the largest, the English corpus as second, the Chinese corpus as third and the Korean collection as the smallest. Table 1 also compares the mean number of distinct bigrams per document, showing that this value is clearly larger for the Chinese collection (363.4 bigrams/article), when compared to the Korean (236.2 bigrams/article) or the Japanese corpus (114.5 bigrams/article). For the English collection, the mean number of distinct words per document is 96.6.

When analyzing the number of pertinent documents per topic, only rigid assessments were considered and thus in this paper only "highly relevant" and "relevant" items are seen as being relevant, under the assumption that only highly or relevant items would be useful for all topics. In certain circumstances however we also assumed those records found to be only somewhat pertinent could be of some value. As a result of this rigid judgment system,

[7] See the Web site http://research.nii.ac.jp/ntcir/.

retrieval effectiveness measures depicted in this paper thus show lower performance levels than they would have with more relaxed assessments. We believe however that the conclusions drawn would be similar, whether we used rigid or relax assessments.

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. By contrast, 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. These divergences may have an impact on some of our merging strategies (see Section 3).

```
<TOPIC>
<NUM> 010 </NUM>
<TITLE> Hu Jintao, Visit, Japan, Korea </TITLE>
<DESC> Find articles pertaining to the activities of the Standing Committee member of
   Politburo China and Hu Jintao's visit to Japan or Korea in 1998. </DESC>
<NARR>
   <BACK> Hu Jintao, ranking 5th in the Standing Committee member of Politburo
   China, left Beijing for a visit to Japan and South Korea on April 21st, 1998. It was his
   first diplomatic visit after being elected Vice President of the National Council in
   China in March of 1998. Hu Jintao had a five-day official visit to Japan and visited
   South Korea from the 26th to the 30th. Please query important scheduled activities of
   Hu Jintao's visit. </BACK>
   <REL> Documents about reports of Hu Jintao's visit to Japan or Korea are relevant.
   Comments about this visit from other countries are not relevant. </REL> </NARR>
<CONC> Hu Jintao, Japan, South Korea, Visit, Activities </CONC> </TOPIC>

<TOPIC>
<NUM> 018 </NUM>
<TITLE> Teenager, Social Problem </TITLE>
<DESC> Find articles dealing with a teenage social problem </DESC>
<NARR>
   <BACK> As materialism appears in many aspects of society, many incidents related
   to young teenagers are becoming a major social problem. </BACK>
   <REL> Articles dealing with specific incidents or social problems related to
   teenagers (age 11 to 19) that show a summary or background story of an incident
   (problem) and information on the teenagers are relevant. Articles only addressing
   general criticisms on youth problems are irrelevant. Incidents or social problems
   where teenagers are mentioned but are not the main issue are partially relevant.
   </REL> </NARR>
<CONC> teenager social problem, youth problem, youth, teenager, human traffic,
   runaway, robbery, suicide, sexual abuse </CONC> </TOPIC>
```

Table 2: Examples of two topics included in the NTCIR-4 test-collection

Following the TREC model, the structure of each topic was based on 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 background context for the topics ("<BACK>"), a relevance assessment criteria ("<REL>"), and finally a concept section ("<CONC>" or C) that provides some related terms (see Table 2 for examples). Rather than limiting them to a

narrow subject range, the topics made available were chosen to reflect a variety of information needs (such as “Viagra,” “North Korea, Starvation, Response,” “Nanotechnology, Realization, Research Trends” or “Japan, Amendment, Law, Self-Defense Force”).

1.2 Search models

In order to ensure that useful conclusions would be obtained when handling new test-collections, we considered it important to evaluate retrieval performance under varying conditions. Thus, in order to obtain this broader view, we evaluated a variety of indexing and search models, ranging from very simple binary indexing schemes to more complex vector-processing schemes.

First we considered adopting a binary indexing scheme in which each document (or topic) was represented by a set of keywords, without assigning any weights (IR model denoted “document=bnn, query=bnn” or “bnn-bnn”). Binary logical restrictions may often be too restrictive for document and query indexing, and it is not always clear whether a document should be indexed by a given term (in this study a single word or bigram). Given that a more appropriate answer is neither “yes” nor “no”, but rather something in between, the term weighting should allow for better differentiation of terms and thus increase indexing flexibility. In this vein, we could also assume that a term's occurrence frequency in a document or in a query (denoted tf) can be a useful feature (IR model denoted “nnn-nnn”).

As a third weighting feature, we might consider that those terms occurring very frequently in the collection do not help us discriminate between relevant and non-relevant items. For this reason we could either count their frequency in the collection or more precisely their inverse document frequency (denoted by idf), resulting in larger weights for more specific terms and smaller weights for more frequent ones. It is important to note here that this specificity does not depend on a given term's semantic properties, but rather it is derived from a statistical notion, or as Sparck Jones says, “we think of specificity as a function of term use” [Sparck Jones 1972]. For example, the word “computer” may be viewed as very specific in a legal corpus because this word appears rarely, whereas in a computer science collection, it would be viewed as a broader term, one that might have a variety of meanings.

Moreover, by using cosine normalization whereby each indexing weight could vary within the range of 0 to 1, we could introduce a technique that usually improves retrieval effectiveness (IR model: “ntc-ntc”). See the Appendix for the exact weighting formulations for the IR models used in this paper.

There are also other variants that we might create, especially in cases where the occurrence of a given term in a document is viewed as a rare event. Thus, it may be good practice to give more importance to the first occurrence of a term, as compared to its successive and repeating occurrences. Therefore, the tf component may be computed as the ln(tf) + 1.0 (“ltc”, “lnc” or “ltn”) or as $0.5+0.5\cdot$[tf / max tf in a document] (“atn”). We might also consider that a term's presence in a shorter document represents stronger evidence than it does in a longer document. In order to take document length into account, more complex IR models have been suggested, including the “Lnu” [Buckley et al. 1996] or the “dtu” IR model [Singhal et al. 1999].

In addition to vector-space approaches, we also considered probabilistic IR models, such as the Okapi probabilistic model [Robertson et al. 2000]. We implemented the Prosit model as a

second probabilistic approach (or "deviation from randomness") [Amati and van Rijsbergen 2002; Amati et al. 2003]. As shown in Equation 1, this IR model combines two information measures. The first component measures the informative content (denoted by $Inf^1_{ij}(tf)$) based on the observation that in the document D_i we found *tf* occurrences of the term t_j. The second one measures the risk (denoted by $1\text{-}Prob^2_{ij}(tf)$) of accepting the term t_j as a good descriptor, knowing that in document D_i there are *tf* occurrences of term t_j.

For the first information factor, $Prob^1_{ij}(tf)$ is the probability of observing *tf* occurrences of the term t_j in document D_i by pure chance. If this probability is high, term t_j may correspond to a noncontent-bearing word in the context of the entire collection [Harter 1975]. For the English language, these words generally correspond to determinants such as "the," prepositions like "with" or verb forms like "is" or "have," considered as being of little or not use in describing a document's semantic content. There are also various nouns that may often appear in numerous documents within a particular corpus, such as "computer" and "algorithm," particularly when the articles in which they are found are extracted from computer science literature. On the other hand, if $Prob^1_{ij}(tf)$ is small (or if $-\log_2[Prob^1(tf)]$ is high), the term t_j would provide important information regarding the content of the document D_i. As defined in Equation 2, in our implementation $Prob^1(tf)$ is expressed as a geometric distribution, where $p = 1/(1+\lambda)$. Other stochastic distributions that have been suggested in Amati and van Rijsbergen [2002].

The term $Prob^2_{ij}(tf)$ represents the probability of having $tf+1$ occurrences of the term t_j, knowing that *tf* occurrences of this term have already been found in document D_i. This probability can be evaluated using the Laplace's law of succession as $Prob^2_{ij}(tf) = (tf + 1) / (tf+2) \approx tf / (tf +1)$ [Dodge 2003, p. 227]. This approximation does not however take document length into account and in our experiments we have included it as shown in Equation 3.

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

$$Prob^1(tf) = [1/(1+\lambda_j)] \cdot [\lambda_j / (1+\lambda_j)]^{tf} \quad \text{with } \lambda_j = tc_j / n \quad (2)$$

$$Prob^2_{ij}(tf) = tfn_{ij} / (tfn_{ij} + 1) \quad \text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl) / l_i)] \quad (3)$$

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, *n* the number of documents in the corpus, *mean dl* the mean length of a document and l_i the length of document D_i.

1.3 Indexing

In the previous section, we described how each indexing unit was weighted so that it would reflect its importance in describing the semantic content of a document or a request. This section will explain how such indexing units are extracted from documents and topic formulations.

For the English collection, we used words as indexing units and based the indexing process on the SMART stopword list (571 terms) and stemmer (in this case, the Lovins' stemming algorithm [1968]). When indexing Indo-European languages, it is natural to consider words as indexing units. For several European languages this approach has usually produced the best retrieval effectiveness, as demonstrated by various CLEF evaluation campaigns [Peters et al. 2004; Peters et al. 2005]. In this case, delimiting words within a sentence is a relatively easy task (with some problems because, for example, "IBM360" or "test-suite" can be viewed

as being composed of one or two words). For various languages there is a list of high-frequency or stoplist containing words that are usually found to be irrelevant when describing the semantic content of documents or queries.

Moreover, in order to conflate word variants into the same stem or root, we also need to adapt a stemming algorithm for each European language. To achieve this goal, we define a light stemming procedure by removing only inflectional suffixes used to indicate number (singular vs. plural), gender (feminine, masculine or neutral) or case (nominative, genitive, ablative, locative, etc.) of a given noun or adjective.[8] Based on the CLEF 2001 test-collections, Savoy [2002] demonstrated that we can obtain mean average precision improvements of 10% (English), 15% (Italian), 18% (Spanish), 21% (German) and 24% (French) when applying a light stemmer, compared to a system use no stemming (T queries). With TDN queries, these improvements are less significant, ranging from 4% (English), to 10% (Spanish & German), or 14% (French & Italian).

More sophisticated stemming strategies also suggest removing certain derivational suffixes (e.g., "-ize," "-ably," "-ship" in the English language). The difference in retrieval effectiveness between light and more complex stemming approaches is usually small. In the French language for example, Savoy [2002] showed that improvements of 5% (T queries) to 2% (TDN queries) would be possible when using an extended stemming procedure.

For the Finnish language however, it seems that the design and development of an effective stemming procedure required a more complex morphological analysis, based on a dictionary. Tomlinson [2004] for example found a statistically significant difference of around 13% in favor of a dictionary-based stemmer, when compared with a derivational stemmer ("Snowball" in this case). For the same language, and based on another set of queries, Moulinier & Williams [2005] confirmed this finding. The real stemming problem with the Finnish language is that stems are often modified when suffixes are added. For example, "matto" (carpet in the nominative singular form) becomes "maton" (in the genitive singular form, with "-n" as suffix) or "mattoja" (in the partitive plural form, with "-a" as suffix). Once we removed the corresponding suffixes, we were left with three distinct stems, namely "matto", "mato" and "matoj". Of course irregularities such as these also occur in other languages, usually helping to make the spoken language flow better, such as "submit" and "submission" in English. In Finnish however, these irregularities are more common, and thus they render the conflation of various word forms into the same stem more problematic.

Finally, most European languages manifest other morphological characteristics, with compound word constructions being the most important (e.g., handgun, worldwide). Braschler & Ripplinger [2004] showed that decompounding German words would significantly improve the retrieval performance, resulting in improvements from 16% to 34% for T queries and 9% to 28% for TDN requests.

In order to develop a language-independent indexing strategy, McNamee & Mayfield [2004] suggested using an overlapping *n*-gram approach to defining the indexing units. In this scheme, each sentence is decomposed into sequences of *n* characters. For example, when analyzing the phrase "the white house", the following 4-grams are extracted {"the_", "he_w", "e_wh", "_whi", "whit", "hite", … "hous", "ouse"}. With this type of indexing approach, stopword lists and stemmers adapted for the corresponding language are not required, since during indexing the *n*-grams appearing in all documents (e.g., "with", "have" or very frequent

[8] Such stopword lists and stemmers are available for various European languages at http://www.unine.ch/clef/ or at http://snowball.tartarus.org/.

suffixes like "-ment") will be assigned null or at least insignificant weights. According to McNamee & Mayfield [2004] and based on eight European languages, the most effective *n*-gram decomposition seems to be between 4-grams and 5-grams.

As explained previously, in the Chinese and Japanese languages words are not clearly delimited. We therefore indexed documents written in Asian languages using an overlapping bigram approach, an indexing scheme that was found to be effective for various Chinese collections [Kwok 1999; Luk and Kwok 2002], or during the NTCIR-3 evaluation campaign [Chen and Gey 2003]. There are also other factors involved in our choice of an indexing tool. When considering the Korean language for example, Lee et al. [1999] found more than 80% of Korean nouns were composed of one or two Hangul characters, and for Chinese Sproat [1992] reported a similar finding. When analyzing the mean length of continuous characters found in the Japanese corpus, we found its value to be 2.3 for Kanji characters, with more than 70% of continuous Kanji sequences being composed of one or two characters. When studying the mean length of continuous Hiragana characters, we calculated an average value of 2.1 and for sequences composed of only of Katakana characters, with a mean value of 3.96.

In our experiments, we have adopted an overlapping bigram approach. In this case, the "ABCD EFG" sequence would generate the following bigrams {"AB," "BC," "CD," "EF" and "FG"}. In order to stop bigram generation in our work we generated these overlapping bigrams for Asian characters only, using spaces and other punctuation marks (as collected for each language from their respective encoding). Moreover, we did not split any words written in ASCII characters. In our experiments, the most frequent bigrams were 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. For the Chinese language, we also evaluated the unigram (or character) indexing approach.

Finally, as suggested by Fujii & Croft [1993] or [Chen and Gey 2003], before generating bigrams for the Japanese documents we removed all Hirakana characters, given that these characters are mainly used to write grammatical words (e.g., *doing*, *do*, *in*, *of*), and the inflectional endings of verbs, adjectives and nouns.

For Asian languages, there are of course other indexing strategies that might be used. In this vein, various authors suggest indexing Chinese documents by using words generated by a segmentation procedure (e.g. based on the longest matching principle [Nie and Ren 1999; Foo and Li 2004]). Nie & Ren [1999] however indicated that retrieval performance based on word indexing does not really depend on an accurate word segmentation procedure and this was confirmed by Foo & Li [2004]. They also stated that segmenting a Chinese sentence does effect retrieval performance and recognizing a greater number of 2-character words usually contributes to retrieval enhancement. These authors did not however find a direct relationship between segmentation accuracy and retrieval effectiveness. Moreover, manual segmentation does not always produce better performance when compared to character-based segmentation.

For the Japanese language, when using the NTCIR-3 test-collection and D topics, Chen & Gey [2003] obtained a mean average precision value of 0.2802 when combining overlapping bigrams and characters vs. 0.2758 for a word-based indexing strategy (words were segmented with the *Chasen* morphological analyzer [Matsumoto et al. 1999]). This performance difference is small (1.6%) and seems to indicate that both indexing schemes result in similar retrieval effectiveness.

For the Korean language, Lee & Ahn [1996] also suggested using *n*-gram representation. In fact, even though word boundaries are marked by spaces, this language also uses numerous

suffixes and even prefixes. Compound constructions are also used very frequently, and a morphological analyzer can be used to separate these compound words into simple nouns. Murata et al. [2003] obtained effective retrieval results using this linguistic approach. However Lee et al. [1999] showed that *n*-gram indexing could provide similar and sometimes better retrieval effectiveness when compared to word-based indexing applied in conjunction with a decompounding scheme.

1.4 Evaluation of various IR systems

Having described the various IR models, it would be useful to know how these search strategies will behave when used with the Asian test-collections. In order to measure retrieval performance we have adopted non-interpolated mean average precision (MAP), as computed by TREC_EVAL. To determine whether or not any given search strategy might be better than another, we based our statistical validation on the bootstrap approach [Savoy 1997]. Thus in the tables appearing in this paper, statistically significant differences are shown underlined (two-sided non-parametric bootstrap test, 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) section, second using the descriptive (D) section and third using all topic logical sections (TDNC).

The mean average precision determined by the eleven search models is shown in Table 3 for the English and Korean collections, with the best performance under a given condition shown in bold (these values were used as a baseline for our statistical tests in Tables 3, 4 and 5). For the Japanese language, Table 4 depicts the performance achieved when generating bigrams from both Kanji and Katakana characters (left part), where in this case a bigram may be composed of one Kanji and one Katakana character. As a variant, we generated bigrams for Kanji characters only, with each continuous Katakana character sequence being considered as a single indexing unit or term. Table 5 shows performances achieved for the Chinese corpus, using the unigram (or character) and bigram indexing schemes.

For Korean (right part of Table 3), Japanese (Table 4) and Chinese (Table 5), the best retrieval models seemed to be the Okapi or the "Lnu-ltc" search models. Surprisingly, this data shows that the best retrieval scheme for short queries was not always the same as that for long topics. For example, for long query formulations (TDNC) and for the Korean collection, the Okapi was the best search model while for short queries (T or D) the vector-space "Lnu-ltc" approach provided better performance. Based on our statistical testing, these differences in performance were not always significant (e.g., with the Japanese corpus, differences between the Okapi and "Lnu-ltc" model were only significant for T queries).

	Mean average precision					
	English (word, 58 queries)			Korean (bigram, 57 queries)		
Model	T	D	TDNC	T	D	TDNC
Prosit	0.2977	0.2871	**0.3803**	0.3882	0.3010	0.4630
Okapi-npn	**0.3132**	0.2992	0.3674	0.4033	0.3475	**0.4987**
Lnu-ltc	0.3069	**0.3139**	0.3524	**0.4193**	**0.4001**	0.4857
dtu-dtn	0.2945	0.2945	0.3126	0.3830	0.3773	0.4397
atn-ntc	0.2808	0.2720	0.3417	0.3604	0.3233	0.4202
ltn-ntc	0.2766	0.2908	0.3271	0.3768	0.3494	0.4224
ntc-ntc	0.1975	0.2171	0.2559	0.3245	0.3406	0.4133
ltc-ltc	0.1959	0.2106	0.2798	0.3103	0.3205	0.4342
lnc-ltc	0.2295	0.2421	0.3235	0.3231	0.3233	0.4616
bnn-bnn	0.1562	0.1262	0.0840	0.1944	0.0725	0.0148
nnn-nnn	0.1084	0.1013	0.1178	0.1853	0.1523	0.1711

Table 3. MAP for various IR models, English and Korean monolingual search

	Mean average precision					
	Bigram on Kanji & Katakana			Bigram only on Kanji		
Model	T	D	TDNC	T	D	TDNC
Prosit	0.2637	0.2573	0.3442	0.2734	0.2517	0.3381
Okapi-npn	**0.2873**	**0.2821**	**0.3523**	**0.2972**	**0.2762**	**0.3510**
Lnu-ltc	0.2701	0.2740	0.3448	0.2806	0.2718	0.3397
dtu-dtn	0.2622	0.2640	0.3221	0.2739	0.2670	0.3161
atn-ntc	0.2424	0.2405	0.3303	0.2543	0.2423	0.3191
ltn-ntc	0.2735	0.2678	0.3265	0.2894	0.2730	0.3249
ntc-ntc	0.2104	0.2087	0.2682	0.2166	0.2101	0.2697
ltc-ltc	0.1868	0.1849	0.2596	0.1926	0.1881	0.2548
lnc-ltc	0.1830	0.1835	0.2698	0.1838	0.1809	0.2633
bnn-bnn	0.1743	0.1741	0.1501	0.1703	0.1105	0.0917
nnn-nnn	0.1202	0.1099	0.1348	0.1184	0.0876	0.0931

Table 4. MAP for various IR models, Japanese monolingual (55 queries)

From a general perspective, it is interesting to note that when using a word-based indexing (English collection, Table 3) or the *n*-gram scheme for Asian languages (Tables 3 to 5), the same top IR models always proved to top performers: Okapi, Prosit, "Lnu-ltc" and "dtu-dtn". Thus, using *n*-grams or words to describe the semantic content of a document (or a request) does not result in any real performance differences among search models. This main conclusion has been corroborated by other studies that compare *n*-gram and word-based indexing strategies when analyzing various European languages [McNamee and Mayfield 2004; Savoy 2002]. Table 6 depicts the mean average precision obtained by using the CLEF 2003 test-collection [Peters et al. 2004] for various European languages belonging to different language groups such as the Roman family (French & Spanish), the German family (German & Dutch, evaluations included a decompounding stage), the Slavic group (Russian) or the Uralic language family (Finnish language) [Savoy 2004c]. As shown in this table, the best performing IR models usually incorporate either the Okapi or the Prosit approach, showing that the performance differences between these two and the "Lnu-ltc" and "dtu-dtn" vector-space models are not always statistically significant.

	Mean average precision					
	Character (or unigram)			Bigram		
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 5. MAP for various IR models, Chinese monolingual (59 queries)

	Mean average precision					
Model	French word 52 queries	Spanish word 57 queries	German word 56 queries	Dutch word 56 queries	Finnish 5-gram 45 queries	Russian word 28 queries
Prosit	**0.5201**	0.4723	0.4553	0.4863	**0.4903**	0.3489
Okapi-npn	0.5164	**0.4885**	**0.4693**	**0.4873**	0.4897	0.3458
Lnu-ltc	0.4826	0.4579	0.4544	0.4508	0.4603	**0.3630**
dtu-dtn	0.4658	0.4503	0.4395	0.4378	0.4354	0.3295
atn-ntc	0.4548	0.4404	0.3932	0.4352	0.4856	0.3322
ltn-ntc	0.3901	0.4240	0.3264	0.3951	0.4294	0.3089
ntc-ntc	0.3274	0.2708	0.3264	0.3036	0.3563	0.3014
ltc-ltc	0.3441	0.2974	0.3602	0.3241	0.3772	0.2874
lnc-ltc	0.3798	0.3353	0.3593	0.3315	0.3721	0.2447
bnn-bnn	0.2401	0.2648	0.2331	0.2680	0.2006	0.1523
nnn-nnn	0.1227	0.1984	0.1085	0.1064	0.1483	0.1141

Table 6. MAP for various IR models using the CLEF 2003 test-collection (monolingual search, TD queries)

As described in Section 1.2, these best IR models share three important common aspects. First, when considering a given indexing unit's occurrence frequency in a document (tf component), they tend to attribute more weight to the first occurrence of this term than to other occurrences. Second, the idf component is also included, e.g., when weighting the search term in the Okapi or "Lnu-ltc" models (or for the Prosit approach in the computation of λ_j in Equation 2). Third and contrary to other IR models, these four best-performing search strategies take document length into account by favoring short documents (usually more focused on a narrow subject).

For the English collection, when analyzing the result language by language (left part of Table 3), the best retrieval scheme seemed to be query-dependant and the best retrieval performance for T queries is the Okapi model, "Lnu-ltc" for D queries and Prosit for TDNC. Moreover, performance differences for these three search models were always statistically significant. While for the English language either the Okapi or the Prosit model will provide

the best retrieval performance [Savoy 2004c; 2005], the good performance shown by the "Lnu-ltc" model using the D queries must be viewed as an outlier.

With the Korean corpus (right part of Table 3) and the T queries, the binary indexing scheme ("bnn-bnn") resulted in surprisingly high retrieval performance when compared to the D or TDNC query formulations (0.1944, 0.0725 and 0.0148 respectively).

For the Japanese collection (Table 4), it is not clear whether bigram generation should have been done for both Kanji and Katakana characters (left part) or only for Kanji characters (right part of Table 4). When using title-only queries, the Okapi model provided the best mean average precision of 0.2972 (bigram on Kanji only) compared to 0.2873 when generating bigrams on both Kanji and Katakana. This difference is rather small, and is even smaller in the opposite direction for long queries (0.3510 vs. 0.3523). Based on these results we cannot infer that for the Japanese language one indexing procedure is always significantly better than another.

For the Chinese collection, when comparing character and bigram representations it seemed that longer queries (TDNC) tended to perform better with bigram indexing. For T or D query constructions, the difference between character and bigram indexing usually favored the bigram approach (the "Lnu-ltc" model performance when using T queries must be viewed an exception). The question that then arises is: How can we improve the retrieval effectiveness of these retrieval models? To answer this question, we suggest that a blind query expansion stage be incorporated during the search process (Section 1.5) and then a fusion strategy applied (Section 1.6).

1.5 Blind query expansion

It is known that once a ranked list of retrieved items has been computed, we can automatically expand the original query by including terms that appear frequently in the top retrieved documents. Called blind query expansion or pseudo-relevance feedback, this technique is performed before presenting the final result list to the user. In this study, we adopted Rocchio's approach [Buckley et al. 1996] with $\alpha = 0.75$, $\beta = 0.75$, whereby the system was allowed to add *m* terms extracted from the *k* best-ranked documents from the original search, as depicted in the following formula.

$$Q' = \alpha \cdot Q + \beta \cdot \frac{1}{|k|} \cdot \sum_{i=1}^{k} D_i \qquad (4)$$

in which Q' indicates the expanded query composed of the previous query Q, and of *m* terms extracted from the *k* best-ranked documents D_i assumed to be relevant to the query Q. Of course other relevance feedback strategies have been proposed. For example, Robertson [1990] suggested making a clear distinction between the term selection procedure and the term weighting scheme. In a similar vein, Carpineto et al. [2001] suggested using a theoretic information measure that in this case was the Kullback-Leibler divergence, for both selecting and weighting terms.

To evaluate this proposition, we used the Okapi and the Prosit probabilistic models. Table 7 summarizes the best results achieved for the English and Korean language collections, Table 8 lists the best retrieval effectiveness achieved with the Japanese corpus (and with our two indexing strategies) as does Table 9 for the Chinese collection (character or bigram indexing). In these tables, the rows labeled "Prosit" or "Okapi-npn" (baseline)

indicate mean average precision before applying this blind query expansion procedure. The rows starting with "#doc. / #terms" indicate the number of top-ranked documents and the number of terms used to enlarge the original query, and thus obtain the best retrieval effectiveness. Finally, the rows labeled "& Q expansion" depict the mean average precision achieved after applying the blind query expansion (using the parameter setting specified in the previous row).

	Mean average precision					
	English (word, 58 queries)			Korean (bigram, 57 queries)		
Model	T	D	TDNC	T	D	TDNC
Prosit	0.2977	0.2871	0.3803	0.3882	0.3010	0.4630
#doc. / #terms	10 / 125	10 / 75	5 / 40	5 / 20	3 / 30	10 / 75
& Q expansion	**0.3731**	**0.3513**	**0.3997**	0.4875	0.4257	0.5126
Okapi-npn	0.3132	0.2992	0.3674	0.4033	0.3475	0.4987
#doc. / #terms	10 / 20	10 / 10	10 / 20	10 / 60	5 / 40	10 / 50
& Q expansion	0.3594	0.3181	0.3727	**0.4960**	**0.4441**	**0.5154**

Table 7. MAP with blind query expansion (English and Korean monolingual)

	Mean average precision					
	Bigram on Kanji & Katakana			Bigram only on Kanji		
Model	T	D	TDNC	T	D	TDNC
Prosit	0.2637	0.2573	0.3442	0.2734	0.2517	0.3381
#doc. / #terms	10 / 300	10 / 100	10 / 125	10 / 100	10 / 100	10 / 100
& Q expansion	**0.3396**	**0.3394**	**0.3724**	0.3495	**0.3218**	**0.3678**
Okapi-npn	0.2873	0.2821	0.3523	0.2972	0.2762	0.3510
#doc. / #terms	10 / 15	5 / 100	5 / 75	10 / 15	10 / 30	5 / 20
& Q expansion	0.3259	0.3331	0.3640	**0.3514**	0.3200	0.3561

Table 8. MAP with blind query expansion, Japanese monolingual (55 queries)

	Mean average precision					
	Character (or unigram)			Bigram		
Model	T	D	TDNC	T	D	TDNC
Prosit	0.1452	0.0850	0.1486	0.1658	0.1467	0.2221
#doc. / #terms	10 / 125	10 / 75	3 / 10	10 / 175	10 / 100	5 / 20
& Q expansion	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. / #terms	10 / 10	10 / 10	10 / 60	5 / 125	5 / 100	5 / 60
& Q expansion	**0.1884**	**0.1407**	**0.2213**	0.2004	0.1805	0.2331

Table 9. MAP with blind query expansion, Chinese monolingual (59 queries)

From the data shown in Tables 7 to 9, we could infer that the blind query expansion technique improved mean average precision, and this improvement is usually statistically significant (values underlined in the tables). When comparing both probabilistic models, this strategy seems to perform better with the Prosit than it does with the Okapi model. For some unknown reason, it seems that we must include more terms with the Prosit model than with the Okapi approach. In addition, the percentage enhancement is greater for short topics than for longer ones. For example, in the Japanese collection (bigram on both Kanji and Katakana) using 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.

Knowing that such query expansion may decrease the retrieval effectiveness for some queries, several variants are proposed. For example, Grunfeld et al. [2003] suggested using the Web to find additional search terms, while Luk & Wong [2004] suggested various term weighting schemes, depending on the term's occurrence in the collection.

1.6 Data fusion

As an additional strategy that would enhance retrieval effectiveness, we considered adopting a data fusion approach that combines two or more result lists provided by different search models.

When adopting this strategy, we assume that different indexing and search models would retrieve different pertinent and non-relevant items, and thus combining the different search models would improve retrieval effectiveness. More precisely, when combining different indexing schemes we would expect to improve recall, due to the fact that different document representations might retrieve different pertinent items [Vogt and Cottrell 1999]. On the other hand, when combining different search schemes, we assume that these various IR strategies are more likely to rank the same relevant items higher on the list than they would the same non-relevant documents (viewed as outliers). Thus, combining them could improve retrieval effectiveness by ranking pertinent documents higher and ranking non-relevant items lower. In this study, we hope to enhance retrieval performance by making use of this second characteristic, while for the Chinese language our assumption would be that character and bigram indexing schemes are distinct and independent sources of evidence regarding the content of documents. For this language only, we expect to improve recall due to the first effect described above.

As a first data fusion strategy, we considered the round-robin (denoted "RR") approach whereby in turn we took one document from all individual lists and removed duplicates, keeping the most highly ranked instances. Various other data fusion operators have been suggested [Fox and Shaw 1994], however the simple linear combination (denoted "SumRSV") usually seemed to provide the best performance [Savoy 2004a; Fox and Shaw 1994]. In this case, for any given set of result lists, this combined operator was defined as $SumRSV = SUM (\alpha_i \cdot RSV_k)$, in which RSV_k denoted the retrieval status value (or document score) of document D_k in the *i*th result list. Finally the value of α_i (set to 1 for all result lists in our experiments) may be used to reflect differences in retrieval performance among the various IR models.

Given that document scores cannot usually be directly compared, as a third data fusion strategy we normalized document scores within each collection through dividing them by the maximum score (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 5.

$$NormRSV_k = ((RSV_k - Min^i) / (Max^i - Min^i)) \quad (5)$$

where Min^i (Max^i) denotes the minimal (maximal) RSV value in the *i*th result list.

As a new data fusion strategy, we suggested merging the retrieved documents according to the Z-score, computed for each result list. For the *i*th result list within this scheme 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 then normalized the document score for each document D_k provided by the *i*th result list, as computed using the following formula:

$$\text{Z-score } RSV_k = \alpha_i \cdot [((RSV_k - Mean^i) / Stdev^i) + \delta^i], \quad \delta^i = ((Mean^i - Min^i) / Stdev^i) \quad (6)$$

where the value of δ^i is used to generate only positive values, and α_i (usually fixed at 1) is used to reflect the relative retrieval performance of the *i*th retrieval model. When the coefficients α_i are not all fixed at 1, the data fusion operator is denoted as "Z-scoreW". The use of the Z-score was also suggested for the topic detection and tracking contexts [Leek et al. 2002].

	Mean average precision								
	Chinese (bigram/character) 59 queries			Japanese (55 queries) bigram on Kanji & Katakana			Korean (bigram) 57 queries		
Model	T	D	TDNC	T	D	TDNC	T	D	TDNC
#doc/#term	5 / 30	10 / 100	10 / 60	10 / 200	10 / 75	10 / 350	10 / 100		3 / 30
Prosit	0.2007	0.1987	**0.2450**	0.3388	0.3390	0.3688	0.4868		0.4657
#doc/#term	5 / 100	10 / 100		5 / 10	10 / 150	10 / 150	3 / 30	5 / 20	10 / 40
Okapi-npn	0.1987	0.1758		0.3181	0.3324	0.3624	0.4654	0.4335	**0.5141**
#doc/#term	3 / 75	5 / 125		10 / 350	5 / 75	10 / 200	10 / 300		
Lnu-ltc	0.1824	0.1711		0.2879	0.2884	0.3545	0.4500		
#doc/#term	10 / 40	5 / 60		10 / 350			5 / 15	5 / 10	
ltn-ntc	0.1780	0.1898		0.2786			0.4303	0.3946	
#doc/#term	10 / 10	3 / 10							
Okapi-npn	0.1884	0.1394	<- character indexing						
#doc/#term	3 / 75	3 / 60							
Lnu-ltc	0.1926	0.1592	<- character indexing						
RR	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-scoreW	0.2120	**0.2011**		**0.3513**	**0.3484**	0.3728	0.5078	**0.4471**	0.5058

Table 10. MAP with various data fusion schemes

Table 10 shows the mean average precision (MAP) obtained from the Chinese, Japanese and Korean collections, for each of the T, D and TDNC queries. The top part of this table shows the individual performances of various retrieval models used in our data fusion experiments. For example, for the T queries in Japanese, we combined the Prosit and Okapi probabilistic models with the "Lnu-ltc" and "ltn-ntc" vector-space schemes. For the Chinese language, data fusion experiments also included the Okapi and "Lnu-ltc" models based on character indexing. The round-robin ("RR") scheme shown in this table was intended to serve as a baseline for our statistical testing.

From this data, we can see that combining two or more IR models sometimes improved retrieval effectiveness. Moreover, linear combinations ("SumRSV") usually resulted in good performance, and the Z-score scheme tended to produce the best performance. As shown in Table 10 under the heading "Z-scoreW", 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 in the Korean collection using TDNC queries. In this case, none of the data fusion operators performed significantly better than the round-robin scheme, while the best single retrieval model (Okapi in this case) was revealed to be the best mean average precision (0.5141). It is difficult however to predict exactly which data fusion operator will produce the best results. The Z-score or the weighted Z-score schemes did however seem to produce good results when handling different languages and query formulations. Our experiments also indicated that combining short query resulted in better improvement than did combining longer topics.

2 Bilingual IR

In order to retrieve information written in one of our Far-East languages based on a topic description written in English, in our approach we made use of freely available resources that automatically provide translations into the 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, namely BabelFish, FreeTranslation, InterTran, WorldLingo, EvDict, Babylon, respectively available at the following locations:

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

When translating a topic into another language, we could also consider parallel and/or comparable corpora. Such an approach is based on document level alignments where, in order to find terms statistically related to the target language, documents in various languages are paired according to their similarity [Braschler and Schäuble 2000]. Comparable corpora were not readily available however, so as a partial solution Nie et al. [1999] suggest using their PTMiner system to extract parallel corpora from the Web. Using these Web page collections, sentences from two pages written in two different languages were aligned using a length-based alignment algorithm [Gale and Church 1993] and the system then computed the probabilities of translating one term into another. When using this type of statistical translation model however source quality (e.g., Web sites) and available corpora size are of prime importance [Nie and Simard 2001]. Cultural, thematic and time differences could also play a role in the effectiveness of these approaches [Kwok et al. 2001].

In the absence of an explicit translation tool, Buckley et al. [1998] suggest that words in one language can be viewed as misspelled forms from another language (for example, English topics are viewed as misspelled French expressions). Following this example, Gey [2004] assumes that Chinese topics can be converted into their Japanese equivalents (after carrying out character set conversion) and hopefully some of the resulting search terms would in fact be appropriate Japanese words. Based on the NTCIR-4 test-collection and using TDNC queries, this author obtained a MAP of 0.0893 when searching the Japanese collection for queries written in Chinese. This retrieval performance represented 25.6% of the corresponding monolingual MAP. Of course such bilingual searches would only work when dealing with related languages, such as Italian and French or, in our context Chinese and Japanese.

When using 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 11 shows the mean average precision obtained when translating English topics employing our two MRDs, the four MT systems and the Okapi model. The first row ("Okapi-npn") of this table also contains the retrieval performances of manually translated topics that will be used as a baseline. The symbol "n/a" in Table 11 represents missing entries, indicating that the translation devices were not able to provide a translation for each language.

	Mean average precision								
	Chinese (bigram) 59 queries			Japanese (55 queries) bigram on Kanji & Katakana			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	0.0795	0.0749	0.1111	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	WorldLingo / EvDict			WorldLingo / Babylon 1			WorldLingo / 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 11. MAP for various query translation approaches (Okapi model)

Based on the T queries and the best single query translation resource, the performance level achieved was only 45.2% of a monolingual search for the Chinese language (0.0795 vs. 0.1755), 67.9% for the Japanese (0.1952 vs. 0.2873) or 46% for the Korean language (0.1855 vs. 0.4033). Machine translation systems thus resulted in generally poor performance levels. Moreover, the differences in mean average precision were always statistically significant and favored manual topic translation approaches. When compared to our previous work with European languages [Savoy 2004c], the differences are clearly larger. For example, during the CLEF 2003 evaluation campaign and using the FreeTranslation MT system, we obtained 82.7% of the performance level achieved by a monolingual search for the French language (0.4270 vs. 0.5164), 80.6% for the Spanish language (0.3997 vs. 0.4885) or 77.4% for the Italian language (0.3777 vs. 0.4880) (evaluation based on English queries). Using the same MT system (FreeTranslation in this case), this query translation comparison reveals that automatic translation from English to other Indo-European languages seems more effective than translating from English into Asian languages.

Moreover, the evaluation performances depicted in Table 11 show that machine translation software tends to produce better query translation than dictionary-based approaches (namely "Babylon" or "EvDict"). Thus automatic query translation operating within a context (the topic formulation in this case) may reduce translation ambiguity. A query-by-query analysis

reveals however that in these experiments the main underlying translation problem is related to the presence of proper nouns (e.g. "Carter", "Torrijos"), geographical terms (e.g., "South Korean") or other proper names (e.g., "Viagra"). From inspecting the Korean queries, we found that these proper nouns were usually not translated and were written in the Latin alphabet in the automatically translated queries (with some exceptions, i.e. "Michael"). Even though the machine usually returned a translation when these terms were translating into Japanese, this suggested translation usually differed from the term used by humans (e.g., "South Korean", while "Apple Computer" seemed to be correctly translated). Moreover, there was no correlation between the performance of translated queries in Japanese and Korean. For example, the machine-based translated Query #7 ("Carter-Torrijos Treaty") performs reasonably well in Korean (0.9188 (bilingual search) vs. 0.9733 (monolingual)) but for the Japanese corpus its performance was better than the monolingual run (0.6847 (bilingual search) vs. 0.3651 (monolingual)). This analysis tends to indicate that we need to consider introducing a supplementary stage during which the Web can be used to provide translation or at least useful related keywords when handling English proper nouns. Kwok et al. [2004] were able to improve the English to Korean search by 18% when using a technique such as this. Chen and Gey [2003] suggested a similar approach for cases when untranslated English words (mainly proper nouns) were found. These terms were submitted to Yahoo!Chinese (or Yahoo!Japan) and the first 200 entries were then downloaded and segmented into words. After this step, they extracted from each line containing the specific English word the five Chinese words immediately to the left and to the right of the English word and included them in the translated topic (assigning a weight $1/k$, with $k = 1$ to 5, to represent the distance between the Chinese and the English words).

Looking at the results language by language, it seems that the BabelFish MT system tends to produce the best translated topics for the Japanese and Korean languages, and both BabelFish and WorldLingo MT for the Chinese. In order to improve retrieval performance, we developed three possible strategies. First, we concatenated the output of two translation tools into a single query. For the Chinese language, we combined the translations given by WorldLingo with those of "EvDict", for the Japanese language we concatenated the translation provided by WorldLingo with those "Babylon 1", and for Korean, WorldLingo and BabelFish. As shown in the last two rows of Table 11, this combined translation strategy seemed to enhance retrieval effectiveness for the Chinese and Japanese languages, but not for Korean.

	Mean average precision								
	Chinese (bigram) 59 queries			Japanese (55 queries) bigram on Kanji & Katakana			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	5 / 60	5 / 60	5 / 75	10 / 75	5 / 75	5 / 75	5 / 75	10 / 200	5 / 60
& Q exp.	0.1039	0.1003	0.1290	**0.2733**	0.2185	0.2669	**0.2397**	**0.2139**	0.2882
Prosit	0.0817	0.0728	0.1133	0.1973	0.1897	0.2508	0.1721	0.1475	0.2409
#doc/#term	5 / 40	10 / 125	5 / 60	10 / 200	10 / 100	10 / 200	10 / 125	10 / 125	10 / 100
& Q exp.	**0.1213**	**0.1057**	**0.1644**	0.2556	**0.2600**	**0.3065**	0.2326	0.2098	**0.2968**

Table 12. MAP for blind query expansion on translated queries (Okapi or Prosit)

In a second attempt to improve performance, we applied a blind query expansion to the combined translated topics. As shown in Table 12, this technique clearly enhanced retrieval

effectiveness when we used the Okapi or Prosit probabilistic models. As for monolingual IR (see Tables 7 to 9) and for the Chinese and Japanese collections, 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 combined 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, blind query feedback seemed to be a very attractive strategy for enhancing retrieval effectiveness.

As a third strategy for enhancing retrieval effectiveness, we might consider adopting a data fusion approach that combines two or more result lists provided by different search models (as shown with the monolingual search, see Section 1.6).

As an additional strategy, it could be useful to know or predict when a given translation is good or when a given search might produce a proper response. In this vein, Kishida et al. [2004b] suggest using a linear regression model to predict the average precision of the current query, based on both manual evaluations of translation quality for the current query and the underlying topic difficulty. Using the 55 queries written in Japanese, together with their machine-based translations from Korean, Chinese and English, these authors found that the 64% variability in average performance was due to both translation quality and also intrinsic query difficulties. In a related paper however, Cronen-Townsend et al. [2002] showed that in monolingual IR a query's idf average value might be able to adequately predict its retrieval effectiveness or intrinsic difficulty. Based on such findings, it could be worthwhile to combine various translations on a per query basis or to select the most appropriate parameters when expanding the original query, also on a per query basis.

3 Multilingual Information Retrieval

In this section, we will investigate situations in which users write a topic in English in order to retrieve relevant documents in English, Chinese and Japanese (CJE) or in English, Chinese, Japanese and Korean (CJKE). To deal with this multi-language hurdle we based our approach on bilingual IR systems, as described in the previous section. Thus, the various collections were indexed separately and once the original requests were received, they were translated into different languages and submitted to the various collections or search engines. As a response, a ranked list of retrieved items was returned from each collection. From these lists we needed to produce a unique ranked result list, using the merging strategy described further on in this section. Moreover, in our multilingual experiments only one search engine would be available, a common situation occurring in digital libraries or in other office environments. Using a good general search engine however we wanted however to compare our various merging strategies, thus we selected the Prosit model. Based on the same test-collection, Savoy [2004b] evaluated various multilingual merging strategies by using a variety of search engines.

Other search strategies have of course been suggested for handling multilingual collections. For example, as an alternative to the query translation approach, we might translate all documents into a single common language [Braschler and Peters 2004; Chen and Gey 2004]. In this case, we might form a huge unique collection with all available documents, and since the search would be performed by comparing with a single collection, no merging procedure would be required. As shown in the CLEF 2003 evaluation campaign, such an indexing and search strategy usually provides very good retrieval effectiveness. The document translation

approach does however require more computational effort and if we allow users to write their queries in *k* languages, we need to translate each document into *k-1* other languages.

In our experiments, we adopted a query translation strategy and then after performing a search on each language, we merged the different result lists. The top part of Table 13 illustrates a merging problem in which a query has been sent to three collections. In response three result lists were received and thus we needed to merge the retrieved items in order to form a unique ranked list, one that reflected the degree of pertinence of each item within the request.

As a first merging approach, we considered the round-robin method [Voorhees et al. 1995], whereby we took one document in turn from all individual lists. In this case, we might assume that each collection (or language in this study) contains approximately the same number of pertinent items and that the distribution of relevant documents is similar across the result lists. Under these hypotheses, the rank of the retrieved documents would be the key feature in generating the final unique result list presented to the user.

As a second approach, and in order to account for the document score computed for each retrieved item (denoted as RSV_k for document D_k), we might formulate the hypothesis that each collection could be searched by the same or a very similar search engine, and that the similarity values would be therefore directly comparable. Such a strategy, called raw-score merging, produced a final list sorted by document score, as computed by each collection. Since we used the same retrieval model (Prosit) for searching within all collections separately, we could expect resulting document scores to be more comparable and thus the document score could be used to sort the retrieved items. However, document scores were not always comparable and this merging strategy favored documents having a high retrieval status value, coming from the Japanese or Korean corpus as illustrated in Table 13.

Japanese collection			Chinese collection			Korean collection		
rank	document	RSV	rank	document	RSV	rank	document	RSV
1	JP015	90	1	ZH167	0.75	1	KR785	60
2	JP256	88	2	ZH572	0.45	2	KR178	54
3	JP678	50	3	ZH719	0.39	3	KR710	51
4	JP961	45	4	ZH739	0.38	4	KR389	30
5	JP178	44	5	ZH078	0.35	5	KR781	29
…	…	…	…	…	…	…	…	…
Round-robin			Raw-score			MaxRSV		
1	JP015		1	JP015	90	1	JP015	1.00
2	ZH167		2	JP256	88	2	ZH167	1.00
3	KR785		3	KR785	60	3	KR785	1.00
4	JP256		4	KR178	54	4	JP256	0.98
5	ZH572		5	KR710	51	5	KR178	0.90
6	KR178		6	JP678	50	6	KR710	0.85
7	JP678		7	JP961	45	7	ZH572	0.60

Table 13. Example of three merging strategies

As demonstrated by Dumais [1994] however, collection-dependent statistics represented by document or query weights may vary widely among collections, and this phenomenon may therefore invalidate the raw-score merging hypothesis, even when the same search engine is used. Thus, as a third scheme, we could normalize the RSV_k by using the retrieved record

document score listed in the first position ("MaxRSV") or by using Equation 5 ("NormRSV"). Under these merging strategies, we assume that document scores computed by search engines working with different corpora are not comparable. Therefore, these document scores must be normalized before using them as a key to sort the retrieved items. As depicted in Table 13, such a merging strategy would account for the difference between a given document score and the document score for the first retrieved item provided by the same collection. In our example, the difference between the first and the third item in the Korean collection is relatively small compared to the difference between the first and the second document in the Chinese collection. Therefore, the third document of the Korean corpus "KR710" must appear before the second document extracted form the Chinese corpus "ZH572".

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. A merging strategy such as this exploits the fact that the Japanese and Korean corpora possess more articles than the English or the Chinese collections (see Table 1). We may thus assume that the Japanese or Korean corpus will contain more pertinent information than the English or Chinese collection.

As a sixth merging approach, we could use our Z-score model (see Section 1.6 & Equation 6) to define a comparable document score across the collections. This merging strategy would exploit the fact that the top-ranked retrieved and pertinent items usually provide a much greater RSV values than do the others, and such documents must be presented to the user. Manmatha et al. [2001] propose a similar idea when modeling the document score distribution in the form of a mixture model. On the other hand, when the document scores from a given result list are all more or less the same, we must consider that such a distribution contains a very large number of irrelevant documents.

In this merging strategy, we may also consider that each collection might have different numbers of pertinent items or that each collection is searched by IR models having different mean retrieval performances. To reflect this bias when using a given collection or search engine, we could multiply each normalized document score by a corresponding weight. Using this idea, under the label "Z-scoreW" we assigned a weight of 1.2 to the Japanese and Korean result lists and 1.0 to the English and Chinese runs. In this study, we increased the weight attached to the Japanese and Korean language due to the fact that these collections contained more documents, and hopefully more relevant documents.

Finally, we could use logistic regression to predict the probability of a binary outcome variable, according to a set of explanatory variables [Le Calvé and Savoy 2000]. In our current case, we predicted the probability that document D_k would be relevant, given both the logarithm of its rank (indicated by $\ln(rank_k)$) and the original document score RSV_k as indicated in Equation 7. Based on these estimated relevance probabilities (computed independently for each language using S+ software), we sorted the records retrieved from separate collections in order to obtain a single ranked list.

$$\text{Prob}\,[D_k \text{ is rel} \mid rank_k,\ rsv_k] = \frac{e^{\alpha+\beta_1\cdot\ln(rank_k)+\beta_2\cdot rsv_k}}{1 + e^{\alpha+\beta_1\cdot\ln(rank_k)+\beta_2\cdot rsv_k}} \quad (7)$$

In order to estimate the underlying parameters however this approach requires that a training set be developed. In our evaluations we did this by using the leaving-one-out approach to produce an unbiased estimation of the real performance. In this case, the training set was made up of all queries except one, and this last request was used to compute the

average precision for this single query. Finally, we iterated them over the query samples, generating 60 different training sets (composed of 59 queries) and 60 query evaluations from which a mean average precision could be computed.

Table 14 shows the retrieval effectiveness of the various merging strategies when English queries were automatically translated. The top part of this table shows the mean average precision obtained independently for each language and using the Prosit search model along with query expansion (number of top-ranked documents / number of additional search terms). The middle part depicts the mean average precision when searching the Chinese, Japanese and English collections (CJE) while the bottom part also includes the Korean language (CJKE). In this table and for both multilingual environments, the round-robin merging strategy served as a baseline upon which statistical tests could be performed. Finally in Table 15, we evaluated multilingual runs using manually translated topics. As depicted in Tables 14 and 15, when comparing differences in retrieval effectiveness we could then estimate any retrieval effectiveness decreases that were due to automatic query translation strategies. Moreover, the experiments shown in Table 15 may be used to confirm the findings found in Table 14.

Table 14 data indicates that only a few runs produced retrieval effectiveness that might be viewed as statistically superior to that of the round-robin baseline. When considering manually translated queries as shown in Table 15, more merging strategies resulted in significantly better performance than did the round-robin scheme.

	Mean average precision		
	T	D	TDNC
English (on 58 queries)	Prosit 10/125	Prosit 10/75	Prosit 5/40
	0.3731	0.3513	0.3997
Chinese (on 59 queries)	Prosit 5/40	Prosit 10/125	Prosit 5/60
Lingo & Ed	0.1213	0.1057	0.1644
Japanese (on 55 queries)	Prosit 10/200	Prosit 10/100	Prosit 10/200
Lingo & Babylon 1	0.2556	0.2600	0.3065
Korean (on 57 queries)	Prosit 10/125	Prosit 10/125	Prosit 10/100
Lingo & BabelFish	0.2326	0.2098	0.2968
Merging strategy on CJE			
Round-robin (baseline)	0.1591	0.1554	0.2040
Raw-score	0.1573	0.1467	0.1914
MaxRSV	0.1671	0.1614	0.2072
NormRSV (Eq. 5)	0.1660	0.1646	0.2129
Biased round-robin (J=2)	0.1657	0.1632	0.2116
Z-score (Eq. 6)	0.1625	0.1613	0.2096
Z-scoreW (Eq. 6) (J=1.2)	0.1673	0.1662	0.2156
Logistic regression	**0.1978**	**0.1917**	**0.2363**
Merging strategy on CJKE			
Round-robin (baseline)	0.1394	0.1343	0.1870
Raw-score	0.1381	0.1292	0.1740
MaxRSV	0.1354	0.1296	0.1718
NormRSV (Eq. 5)	0.1407	0.1379	0.1871
Biased round-robin (J=K=2)	0.1412	0.1358	0.1885
Z-score (Eq. 6)	0.1406	0.1397	0.1941
Z-scoreW (Eq. 6) (J=K=1.2)	0.1430	0.1421	0.1970
Logistic regression	**0.1676**	**0.1630**	**0.2187**

Table 14. MAP of various merging strategies for CJE collection (medium) and CJKE collection (bottom) with automatic query translation

As a first approach, both simple and normalized merging schemes ("MaxRSV" or "NormRSV") provided reasonable performance levels, with the "NormRSV" merging scheme being slightly better. In our experiments, while the raw-score approach's retrieval effectiveness was not very good, performance decreases were not usually statistically significant when compared to the round-robin scheme (except for manually translated queries and CJKE search as shown in the bottom part of Table 15). Our biased round-robin scheme seems to perform better when compared to the simple round-robin version, yet it is difficult a priori to know whether any given corpus will actually contain more relevant items than another. In this study, we assumed that the number of documents in a given collection was correlated with the number of relevant items contained in this corpus. Both the Z-score and the weighted Z-score (with $\alpha = 1$ for the English and Chinese corpora and 1.2 for both the Japanese and Korean languages) usually achieved better performance levels than did the round-robin approach (performance differences were not however always statistically significant, at least in Table 14). For all multilingual searches, our logistic merging scheme produced the best mean average precision and was always statistically superior to the round-robin approach. As a second best approach, Tables 14 and 15 indicate that our weighted Z-score merging scheme always produced the best second retrieval performance.

	Mean average precision		
	T	D	TDNC
English (on 58 queries)	Prosit 10/125 0.3731	Prosit 10/75 0.3513	Prosit 5/40 0.3997
Chinese (on 59 queries)	Prosit 10/175 0.2140	Prosit 10/100 0.1987	Prosit 5/20 0.2507
Japanese (on 55 queries)	Prosit 10/300 0.3396	Prosit 10/100 0.3394	Prosit 10/125 0.3724
Korean (on 57 queries)	Prosit 5/20 0.4875	Prosit 3/30 0.4257	Prosit 10/75 0.5126
Merging strategy on CJE			
Round-robin (baseline)	0.2230	0.2139	0.2505
Raw-score	0.2035	0.1981	0.2364
MaxRSV	0.2222	0.2180	0.2541
NormRSV (Eq. 5)	0.2281	0.2195	0.2560
Biased round-robin (J=2)	0.2345	0.2260	0.2624
Z-score (Eq. 6)	0.2293	0.2243	0.2620
Z-scoreW (Eq. 6) (J=1.2)	0.2351	0.2320	0.2716
Logistic regression	**0.2505**	**0.2396**	**0.2827**
Merging strategy on CJKE			
Round-robin (baseline)	0.2305	0.2157	0.2636
Raw-score	0.1913	0.1879	0.2430
MaxRSV	0.2210	0.2038	0.2645
NormRSV (Eq. 5)	0.2305	0.2139	0.2674
Biased round-robin (J=K=2)	0.2393	0.2234	0.2734
Z-score (Eq. 6)	0.2395	0.2273	0.2770
Z-scoreW (Eq. 6) (J=K=1.2)	0.2462	0.2361	0.2866
Logistic regression	**0.2549**	**0.2422**	**0.2981**

Table 15. MAP of various merging strategies applied to CJE collection (medium) and CJKE collection (bottom) with manually query translation

Finally, performance differences between manually and automatically translated queries were relatively significant. For the CJE multilingual retrieval and T queries, the best automatic run had a mean average precision rate of 0.1978 compared to 0.2505 (or a 21% difference in relative performance). When compared with the CJKE multilingual search and T queries, the difference was greater (0.1676 vs. 0.2549, or 34.2%).

In addition to retrieval effectiveness, it would be worthwhile to obtain an overview of computational efforts required to build and search these test-collections. The top part of Table 16 lists the size of each collection in terms of storage space requirements and number of documents. The "# postings" row indicates the number of terms (words for the English corpus, bigrams for the three Asian languages) in the inverted file. The next row shows the inverted file size and the following row depicts the time (user CPU time + system CPU time) needed to build this inverted file. The other rows show the average query size and search time (in seconds) required for both short (T) and long (TDNC) queries (measured without blind query expansion). To implement and evaluate these various search models, we used an Intel Pentium III/600 (memory: 1 GB, swap: 2 GB, disk: 6 x 35 GB).

	English	Chinese	Japanese	Korean
# 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 queries				
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 queries				
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 16. Inverted file and search statistics (NTCIR-4 test-collection)

Conclusion

Successful access to multilingual document collections requires an effective monolingual indexing and search system, a combined query translation approach and a simple but efficient merging strategy [Braschler and Peters 2004; Chen and Gey 2004; Savoy 2004a]. Using this blueprint derived during the latest CLEF evaluation campaigns, we effectively applied it to the three Far-East Asian languages. Thus, as a result of our evaluations when indexing Asian languages based on bigrams, the "Lnu-ltc" vector-space or the Okapi probabilistic IR models (see Tables 3 to 5) achieve the best retrieval performance levels. Blind-query expansion has proven to be a worthwhile approach, especially when processing short queries and using the Prosit IR model (see Tables 7 though 9). In order to further improve retrieval effectiveness, a data fusion approach could be considered, although this technique would require additional computational resources (see Table 10).

Based on our analysis of bilingual search performances, our results conflicted with those found for certain European languages [Savoy 2004a; 2004c], especially given the number and questionable quality of freely available translation resources. Thus, when compared with corresponding monolingual searches in which we translated user information needs from English into Chinese, Japanese or Korean language, overall retrieval effectiveness decreases more than 30% for the Japanese, and more than 50% for the Chinese and Korean languages (see Table 11). To improve this poor performance, we could concatenate two (or more) translations (see the last two rows of Table 11), employ a blind query expansion approach (see Table 12), and a data fusion approach.

When evaluating various merging strategies using different query sizes, it appears that when merging ranked lists of retrieved items provided by separate collections, good retrieval effectiveness is obtained with the Z-score merging procedure (around 5% better that round-robin approach). When a representative query sample is available however the logistic merging scheme always produces the best retrieval effectiveness (between 10% (CJKE, manual query translation, T queries) to 24% (CJE, automatic query translation, T queries) better that round-robin approach).

Acknowledgments

The author would like to thank the task NTCIR-4 CLIR organizers for their efforts in developing the Asian test-collections and the anonymous referees for their useful suggestions and comments. The author would also like to thank C. Buckley from SabIR for giving us the opportunity to use the SMART system, together with Samir Abdou and Pierre-Yves Berger for their help in translating the English topics. This research was supported by the Swiss National Science Foundation under Grant #20-103420/1.

References

Amati, G. and van Rijsbergen, C.J. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20, 4, 357-389.

Amati, G., Carpineto, C., and Romano, G. 2003. Italian monolingual information retrieval with PROSIT. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, (Eds.), *Advances in Cross-Language Information Retrieval*, Lecture Notes in Computer Science, 2785, (pp. 257-264). Berlin: Springer-Verlag.

Bloomfield, L. 1933. Language. New York, NY: Holt, Rinehart and Winston.

Buckley, C., Singhal, A., Mitra, M. and Salton, G. 1996. New retrieval approaches using SMART. In *Proceedings of TREC-4*, Gaithersburg, MD, November 1995. D.K. Harman (Eds), NIST Special Publication 500-236, 25-48.

Buckley, C., Mitra, M., Waltz, J. and Cardie, C. 1998. Using clustering and superconcepts within SMART. In *Proceedings of TREC-6*, Gaithersburg, MD, November 1997. E.M. Voorhees and D.K. Harman (Eds), NIST Special Publication 500-240, 107-124.

Braschler, M. and Schäuble, P. 2000. Using corpus-based approaches in a system for multilingual information retrieval. *IR Journal*, 3, 3, 273-284.

Braschler, M. and Peters, C. 2004. Cross-language evaluation forum: Objectives, results and achievements. *IR Journal*, 7, 1-2, 7-31.

Braschler, M. and Ripplinger, B. 2004. How effective is stemming and decompounding for German text retrieval? *IR Journal*, 7, 3-4, 291-316.

Carpineto, C., De Mori, R., Romano, G. and Bigi, B. 2001. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19, 1, 1-27.

Chen, A. and Gey, F.C. 2003. Experiments on cross-language and patent retrieval at NTCIR-3 workshop. In *Proceedings of NTCIR-3*, N. Kando (Ed.), Tokyo, 2003.

Chen, A. and Gey, F.C. 2004. Multilingual information retrieval using machine translation, relevance feedback, and decompounding. *IR Journal*, 7, 1-2, 149-182.

Cronen-Townsend, S., Zhou, Y. and Croft, W.B. 2002. Predicting query performance. In *Proceedings of the 25th International Conference on Research and Development in Information Retrieval*. K. Jarvelin, M. Beaulieu, R. Baeza-Yates and S.H. Myaeng (Eds.), New York, NY: The ACM Press, 299-306.

Dumais, S.T. 1994. Latent semantic indexing (LSI) and TREC-2. In *Proceedings TREC-2*, Gaithersburg, MD, September 1993. D.K. Harman (Eds), NIST Special Publication 500-215, 105-115.

Dodge, Y. (Ed.) 2003. The Oxford dictionary of statistical terms. Oxford, UK: Oxford University Press.

Foo, S. and Li, H. 2004. Chinese word segmentation and its effect on information retrieval. *Information Processing & Management*, 40, 1, 161-190.

Fox, E.A. and Shaw, J.A. 1994. Combination of multiple searches. In *Proceedings TREC-2*, Gaithersburg, MD, September 1993. D.K. Harman (Eds), NIST Special Publication 500-215, 243-249.

Fujii, H., and Croft, W.B. 1993. A comparison of indexing techniques for Japanese text retrieval. In *Proceedings of the 16th International Conference on Research and Development in Information Retrieval*. New York, NY: The ACM Press, 237-246.

Gale, W. A. and Church, K. W. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistic*, 19, 1, 75-102.

Gey, F. 2004. Chinese and Korean topic search of Japanese news collections. In *Working Notes of NTCIR-4*, N. Kando (Ed.), Tokyo, June 2004, 214-218.

Grunfeld L., Kwok, K. L., Dinstl, N. and Deng, P. 2004. TREC2003 robust, HARD and QA track experiments using PIRCS. In *Proceedings TREC-12*, Gaithersburg, MD, November 2003. E.M. Voorhees and D.K. Harman (Eds), NIST Special Publication 500-255, 510-521.

Halpern, J. 2002. Lexicon-based orthographic disambiguation in CJK intelligent information retrieval. In *Proceedings of COLING-02: The 3rd Workshop on Asian Language Resources and International Standardization.*

Harter, S.P. 1975. A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature. *Journal of the American Association for Information Science*, 26, 197-216.

Kishida, K., Chen, K.H., Lee, S., Kuriyama, K., Kando, N., Chen, H.H., Myaeng, S.H. and Eguchi, K. 2004a. Overview of CLIR Task at the Forth NTCIR Workshop. In *Working Notes of NTCIR-4*, N. Kando (Ed.), Tokyo, June 2004, 1-59.

Kishida, K., Kuriyama, K., Kando, N. and Eguchi, K. 2004b. Prediction of performance on cross-lingual information retrieval by regression models. In *Working Notes of NTCIR-4*, N. Kando (Ed.), Tokyo, June 2004, 219-224.

Kwok, K.L., Grunfeld L., Dinstl, N. and Chan, M. 2001. TREC-9 Cross-language, Web and question-answering track experiments using PIRCS. In *Proceedings TREC-9*, Gaithersburg, MD, November 2000. E.M. Voorhees and D.K. Harman (Eds), NIST Special Publication 500-249, 417-426.

Kwok, K.L. (1999). Employing multiple representations for Chinese information retrieval. *Journal of the American Society for Information Science*, 50, 8, 709-723.

Kwok, K.L., Dinstl, N. and Choi, S. 2004. NTCIR-4 Chinese, English, Korean cross-language retrieval experiments using PIRCS. In *Working Notes of NTCIR-4*, N. Kando (Ed.), Tokyo, June 2004, 186-192.

Le Calvé, A. and Savoy, J. 2000. Database merging strategy based on logistic regression. *Information Processing & Management*, 36, 3, 341-359.

Lee, J.H. and Ahn, J.S. 1996. Using n-grams for Korean text retrieval. In *Proceedings of the 19th International Conference on the ACM-SIGIR'96*, (pp. 216-224). H.P. Frei, D. Harman, P. Schäuble and R. Wilkinson (Eds), New York, NY: The ACM Press.

Lee, J.J., Cho, H.Y. and Park, H.R. 1999. N-gram-based indexing for Korean text retrieval. *Information Processing & Management*, 35, 4, 427-441.

Leek, T., Schwartz, R. and Srinivasa, S. 2002. Probabilistic approaches to topic detection and tracking. In J. Allan (Ed.), *Topic Detection and Tracking: Event-based Information Organization*, (pp. 67-83). Boston, MA: Kluwer.

Lovins, J.B. 1982. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11, 1, 22-31.

Luk, R.W.P. and Kwok, K.L. 2002. A comparison of Chinese document indexing strategies and retrieval models. *ACM Transactions on Asian Language Information Processing*, 1, 3, 225-268.

Luk, R.W.P. and Wong, K.F. 2004. Pseudo-relevance feedback and title re-ranking for Chinese information retrieval. In *Working Notes of NTCIR-4*, Tokyo, N. Kando (Ed.), June 2004, 206-213.

Lunde, K. 1998. CJKV Information Processing. Chinese, Japanese, Korean & Vietnamese Computing. New York, NY: O'Reilly.

Manmatha, R., Rath, T. and Feng, F. 2001. Modeling score distributions for combining the outputs of search engines. In *Proceedings of the 24th International Conference on the ACM-SIGIR'2001*, (pp. 267-275). D.H. Kraft, W.B. Croft, D.J. Harper and J. Zobel (Eds), New York, NY: The ACM Press.

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H. and Asahara, M. 1999. Japanese morphological analysis system ChaSen. Technical Report NAIST-IS-TR99009, NAIST, available at http://chasen.aist-nara.ac.jp/

McNamee, P., and Mayfield, J. 2004. JHU/APL experiments in tokenization and non-word translation. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, (Eds.), *Comparative Evaluation of Multilingual Information Access Systems*, Lecture Notes in Computer Science, 3237, (pp. 85-97). Berlin: Springer-Verlag.

Moulinier, I., and Williams, K. 2005. Report on Thomson legal and regulatory experiments at CLEF 2004. In C. Peters, P. Clough, J. Gonzalo, G. Jones, M. Kluck, B. Magnini, (Eds.) *Advances in Cross-Language Information Retrieval*. Lecture Notes in Computer Science (to appear). Berlin: Spinger-Verlag.

Murata, M., Ma, Q. and Isahara, H. 2003. Applying multiple characteristics and techniques to obtain high levels of performance in information retrieval. In *Proceedings of NTCIR-3*, N. Kando (Ed.), Tokyo, 2003.

Nie, J.Y. and Ren, F. 1999. Chinese information retrieval: using characters or words? *Information Processing & Management*, 35, 4, 443-462.

Nie, J. Y., Simard, M., Isabelle, P. and Durand, R. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd International Conference of the ACM-SIGIR'99*, (pp. 74-81). M. Hearst, F. Gey and R. Tong (Eds), New York, NY: The ACM Press.

Nie, J. Y. and Simard, M. 2001. Using statistical translation models for bilingual IR. In C. Peters, M. Braschler, J. Gonzalo and M. Kluck (Eds.), *Evaluation of Cross-language Information retrieval Systems*, (pp. 137-150). Berlin: Springer-Verlag.

Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Eds.) 2004. *Advances in Cross-Language Information Retrieval*. Lecture Notes in Computer Science (2785). Berlin: Springer-Verlag.

Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M., Magnini, B. (Eds.) 2005. *Advances in Cross-Language Information Retrieval*. Lecture Notes in Computer Science (to appear). Berlin: Springer-Verlag.

Robertson, S.E. 1990. On term selection for query expansion. *Journal of Documentation*, 46, 4, 359-364.

Robertson, S.E., Walker, S. and Beaulieu, M. 2000. Experimentation as a way of life. *Information Processing & Management*, 36, 1, 95-108.

Savoy, J. 1997. Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33, 4, 495-512.

Savoy, J. 2002. Recherche d'information dans des corpus plurilingues. *Ingénierie des systèmes d'informations*, 7, 1-2, 63-93.

Savoy, J. 2004a. Combining multiple strategies for effective monolingual and cross-lingual retrieval. *IR Journal*, 7, 1-2, 121-148.

Savoy, J. 2004b. Report on CLIR task for the NTCIR-4 evaluation campaign. In *Working Notes of NTCIR-4*, N. Kando (Ed.), Tokyo, June 2004, 178-185.

Savoy, J. 2004c. Report on CLEF-2003 monolingual tracks: Fusion of probabilistic models for effective monolingual retrieval.. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, (Eds.), *Comparative Evaluation of Multilingual Information Access Systems*, Lecture Notes in Computer Science, 3237, (pp. 322-336). Berlin: Springer-Verlag.

Savoy, J. 2005. Data fusion for effective European monolingual information retrieval. In C. Peters, P. Clough, J. Gonzalo, G. Jones, M. Kluck, B. Magnini, (Eds.) *Advances in Cross-Language Information Retrieval*. Lecture Notes in Computer Science (to appear). Berlin: Spinger-Verlag.

Singhal, A., Choi, J., Hindle, D., Lewis, D.D. and Pereira, F. 1999. AT&T at TREC-7. In *Proceedings of TREC-7*, Gaithersburg, MD, November 1998. E.M. Voorhees and D.K. Harman (Eds), NIST Special Publication 500-242, 239-251.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 1, 11-21.

Sproat, R. 1992. *Morphology and Computation*. Cambridge, MA: The MIT Press.

Tomlinson, S. 2004. Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, (Eds.), *Comparative Evaluation of Multilingual Information Access Systems*, Lecture Notes in Computer Science, 3237, (pp. 286-300). Berlin: Springer-Verlag.

Vogt, C.C. and Cottrell, G.W. 1999. Fusion via a linear combination of scores. *IR Journal*, 1, 3, 151-173.

Voorhees, E.M., Gupta, N.K. and Johnson-Laird, B. 1995. The collection fusion problem. In *Proceedings of TREC-3*, Gaithersburg, MD, November 1994. D.K. Harman (Eds), NIST Special Publication 500-225, 95-104.

Appendix

In Table A.1, w_{ij} represents the indexing weight assigned to term t_j in document D_i. To define this value, we used *n* to indicate the number of documents in the collection and nt_i the number of distinct indexing units (bigrams or terms) included in the representation of D_i. We assigned values to the constant *b* as follows: 0.5 for both the Chinese and Japanese corpora, 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, we assigned $c = 2$ for the Japanese and Korean corpus, $c = 1$ for the English and $c = 1.5$ for the Chinese. These values were chosen because they usually result in better retrieval performance levels. Finally, 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	$w_{ij} = \frac{\ln(tf_{ij})+1}{\sqrt{\sum_{k=1}^{t}(\ln(tf_{ik})+1)^2}}$	ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t}(tf_{ik} \cdot idf_k)^2}}$
atn	$w_{ij} = idf_j \cdot [0.5+0.5 \cdot tf_{ij}/maxtf_{i.}]$			dtn	$w_{ij} = [\ln(\ln(tf_{ij})+1) + 1] \cdot idf_j$
Lnu	$w_{ij} = \frac{\left((1+\ln(tf_{ij}))/(\ln(\text{mean tf})+1)\right)}{(1-\text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$			ltc	$w_{ij} = \frac{(\ln(tf_{ij})+1) \cdot idf_j}{\sqrt{\sum_{k=1}^{t}((\ln(tf_{ik})+1) \cdot idf_k)^2}}$
Okapi	$w_{ij} = \frac{((k_1+1) \cdot tf_{ij})}{(K+tf_{ij})}$			dtu	$w_{ij} = \frac{(\ln(\ln(tf_{ij})+1)+1) \cdot idf_j}{(1-\text{slope}) \cdot \text{pivot} + (\text{slope} \cdot nt_i)}$

Table A.1. Weighting schemes