Berkeley at NTCIR-2: Chinese, Japanese, and English IR Experiments

Aitao Chen,* Fredric C. Gey† and Hailing Jiang*
*School of Information Management and Systems
†UC Data Archive & Technical Assistance (UC DATA)
University of California at Berkeley, CA 94720, USA
{aitao,hjiang1}@sims.berkeley.edu, gey@ucdata.berkeley.edu

Abstract

This paper reports on the work of Berkeley group at the second NTCIR workshop on Japanese & English IR and Chinese IR. A number of runs were submitted on all subtasks in the two main tasks. Our main focus on the Japanese monolingual subtask was on comparing the retrieval effectiveness of different segmentation methods. The experimental results show the bigram indexing outperformed the word-based indexing in Japanese monolingual retrieval. The bigram indexing was also highly effective in Chinese monolingual retrieval. This paper presents an alternative segmentation method that breaks text into one-character terms and two-character terms that do not overlap with each other, which overcomes the disadvantage of producing large index files by overlapping bigram indexing. This paper describes a technique for building bilingual word lexicons from parallel text by sentence alignment and word association. A purely rank-based document pooling strategy is presented for combining monolingual retrieval results in multilingual retrieval.
Keywords: *Japanese IR, Chinese IR, Cross-language IR, word segmentation, word alignment*

1 Introduction

At the second NTCIR workshop, Berkeley participated in the *Japanese & English Information Retrieval Task* and the *Chinese Information Retrieval Task*. The Japanese & English IR task consists of six subtasks, covering Japanese and English monolingual retrieval, cross-language retrieval between Japanese and English, and multilingual retrieval from mixed Japanese and English document collection. The Chinese IR task has two subtasks involving Chinese monolingual retrieval and English-to-Chinese cross-language retrieval. We submitted a number of retrieval runs on all subtasks in the two main tasks. For all of our retrieval runs, the same retrieval algorithm was used. Our main focus on the Japanese monolingual retrieval (i.e. the J-J subtask) was on comparing the retrieval effectiveness of different segmentation methods. The availability of the pre-segmented test collection enables us to compare retrieval effectiveness of word-based indexing and bigram indexing in Japanese monolingual retrieval while holding other factors constant. A disadvantage of using overlapping bigram indexing is that the index files are much larger than the word index files generated from the same collection. We presented an alternative indexing technique that breaks text into unigrams and bigrams that do not overlap, which overcomes the disadvantage of producing large index files by overlapping bigram indexing. For the cross-language retrieval between Japanese and English, we created a bilingual keyword lexicon and a bilingual word lexicon from the documents with English translations in the NTCIR-1 collection. These two lexicons were used to translate the Japanese topics into English and vice versa. For the multilingual retrieval subtasks, we presented a simple rank-based document pooling strategy for combining monolingual retrieval results to produce a single ranked list of documents.

This work builds on our earlier work on full-text monolingual and cross-language information retrieval [6, 4, 3, 11], word segmentation [2], and word alignment [5]. In next section, we will describe the document ranking formula used in all of our retrieval runs.

2 Document Ranking

The document ranking formula we used in all of our retrieval runs was Berkeley's TREC-2 formula [6]. The ad hoc retrieval results on the TREC test collections have shown that the formula is robust for long queries and manually reformulated queries, and the results of applying the same formula to the TREC-5 Chinese collection further demonstrated the robustness of the formula [12]. The logodds of relevance of document D to query Q is given by

$$\begin{aligned}\log O(R|D,Q) &= log\frac{P(R|D,Q)}{P(\overline{R}|D,Q)} \\ &= -3.51 + 37.4 * X_1 + 0.330 * X_2 + \\ &\quad -0.1937 * X_3 + 0.0929 * X_4\end{aligned}$$

where $P(R|D,Q)$ is the probability of relevance of document D with respect to query Q, $P(\overline{R}|D,Q)$ is the probability of irrelevance of document D with respect to query Q. The four composite variables X_1, X_2, X_3, and X_4 are defined as follows:

$$X_1 = \frac{1}{\sqrt{N}+1} \sum_{i=1}^{N} \frac{qtf_i}{ql+35}$$

$$X_2 = \frac{1}{\sqrt{N}+1} \sum_{i=1}^{N} \log \frac{dtf_i}{dl+80}$$

$$X_3 = \frac{1}{\sqrt{N}+1} \sum_{i=1}^{N} \log \frac{ctf_i}{cl}$$

$$X_4 = N$$

where N is the number of matching terms between a document and a query, qtf_i is the within-query frequency of the ith matching term, dtf_i is the within-document frequency of the ith matching term, ctf_i is the occurrence frequency in a collection of the ith matching term, ql is query length (number of terms in a query), dl is document length (number of terms in a document), and cl is collection length, i.e. the number of occurrences of all terms in a test collection.

The relevance probability of document D with respect to query Q can be written as follows given the logodds of relevance.

$$P(R|D,Q) = \frac{1}{1+e^{-logO(R|D,Q)}} \quad (1)$$

The documents are ranked in decreasing order by their relevance probability $P(R|D,Q)$ with respect to a query. The ranking formula combines a small set of composite relevance clues which in turn are expressed in primitive relevance clues such as the number of matching terms between a document and a query, the within-document term frequency, the document length, the within-query term frequency, query length, within-collection term frequency, and so on. The coefficients were determined by fitting the logistic regression model to a set of training records using a statistical software package. We refer readers to reference [6] for more details.

3 Chinese IR Task

The Chinese IR task is concerned with Chinese monolingual retrieval and English-to-Chinese cross-language retrieval. The test collection contains 50 topics in Chinese and in English and 132,173 news articles from five newspapers published in Taiwan. The topics consists of *title*, *question*, *narrative*, and *concept* fields. The topics are rich in concept terms. The number of concept terms in topics ranges from 9 to 20 with an average of 15.46 concept terms. We performed two Chinese monolingual retrieval and one English-to-Chinese cross-language retrieval runs.

3.1 C-C Subtask

The focus of this substask is the Chinese monolingual retrieval. Since word boundaries are not marked in Chinese written text, some preprocessing is needed to break Chinese sentences into indexing terms, which can be words, single characters, two-characters, and so on. The process of splitting text into words is called word segmentation. While breaking sentences into words may be necessary for natural language processing and understanding tasks, such as part-of-speech tagging, word sense disambiguation, syntactical parsing, and the like, words are not the only indexing unit for the purpose of information retrieval where the main concern is to retrieve from the document collection those documents that are most likely relevant in response to a user query. In practice, treating two adjacent characters as indexing terms is not only simple to apply, but effective as well. It may not be efficient in space, however it takes no external linguistic resources to index a collection. Thus this approach can be readily applied to documents in any domain. Splitting text into words often needs at least a list of words. The coverage of the word list can have significant effect on the word segmentation results. The proper nouns and domain-specific terminological terms are problematic because often they are missing in dictionaries or word lists used to segment a collection. Our experience with overlapping bigram indexing in both Chinese and Japanese monolingual retrieval is encouraging. Two official runs, named Brkly-CHIR-TI-01 and Brkly-CHIR-LO-01, were submitted for the Chinese monolingual subtask. These two runs differ only in the number of topic fields indexed in the topics. The Brkly-CHIR-TI-01 run indexed the *question* field only in the topic, while the Brkly-CHIR-LO-01 run indexed all four topic fields: the *title*, *question*, *narrative*, and *concepts*. The overlapping bigram indexing was applied to the topics and the document collection. The average precision values over 50 topics are presented in table 1. The overall average precision values of Brkly-CHIR-LO-01 with respect to the relaxed and rigid relevance judgment sets are .7027 and .6073, respectively. This run achieved 99.55% and 100% overall recall with respect to the relaxed and rigit relevance judgement sets. The high average precision may be attributed to the richness in concept terms in the Chinese topics.

3.2 E-C Subtask

The English-to-Chinese IR subtask is about searching English topics against the Chinese document collection. Our approach is to translate the English topics into Chinese by dictionary lookup. Then we performed a monolingual retrieval with the automatic Chinese translations of the English topics.

The topics were processed in three steps to generate the queries before translation. First, the topics were

Run ID	Topic Fields Indexed	Document Fields Indexed	Topic/Document Segmentation Method	Average Precision (Relaxed)	Average Precision (Rigid)
Brkly-CHIR-LO-01	T,Q,N,C	Title and Text	overlapping bigram	0.7027	0.6073
Brkly-CHIR-TI-01	Q	Title and Text	overlapping bigram	0.4758	0.3274

Table 1. This table shows the fields indexed in topics and documents and the segmentation methods used to break documents and topics into words.

tagged using Brill's part-of-speech tagger [1]. Second, noun phrases are extracted from the tagged topics. Third, the single-word terms and phrases are normalized using a morphological analyzer. The following text shows the tagged text of the question field in topic 030.

> To/TO retrieve/VB the/DT reasons/NNS and/CC situations/NNS of/IN widespread/JJ of/IN the/DT infectious/JJ diseases/NNS caused/VBN by/IN El/NNP Nino/NNP ./.

Each word is followed by its part-of-speech tag. The tags NN and NNS represent singular nouns and plural nouns, respectively; NNP represents the proper name, and JJ represents adjective. Then the tagged text is passed to a noun phrase recognizer for noun phrase extraction. The recognizer detects simple noun phrases based on the pattern of the tags. The noun phrase patterns we used to extract noun phrases can be concisely specified in a three-state automaton as shown in Figure 1. The initial state is 0 and the final state is 2. Any words tagged with part-of-speech tags NN, NNS, NNP, NP and NPS are represented by the label NOUN, and words tagged with JJ, JJR, and JJS, which are the positive, comparative and superlative form of an adjective, are represented by the label ADJ. Any sequence of words whose part-of-speech tags completes a path from the initial state to the final state will be extracted as a noun phrase, excluding the single-word nouns. The noun phrases extracted from the above

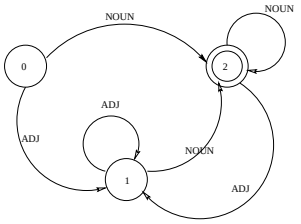

Figure 1. An automaton for simple noun phrase recognition

tagged text are *infectious diseases* and *El Nino*. The words appearing in the stoplist were removed and then the remaining single words and noun phrases are normalized using a morphological analyzer [7], which reduces plural nouns to their singular form and verbs to their base form. Also, all words and phrases are converted to lower case. The normalized single words and the simple noun phrases constitute the English queries before translation.

After the preprocessing of the English topics, each query now consists of single words and noun phrases. We translate each query by looking up every single word and noun phrase in a Chinese-English bilingual dictionary. The bilingual dictionary we used to translate English queries is the Chinese-to-English wordlist (version 2.0) compiled by Linguistic Data Consortium which can be downloaded from from http://morph.ldc.upenn.edu/Projects/Chinese/. The wordlist contains some 128,000 Chinese words, paired with equivalent English words.

For Brkly-ECIR-LO-01 run, a query term (noun phrase or single word) was looked up in the LDC bilingual wordlists. The top two Chinese translation equivalents that occur most frequently in the test document collection were retained as translations for an English term when there are more than two translations for that term. When there was no exact match for a single-word term, that term was not translated. However for no exact match for a noun phrase, we proceeded to match the sub-phrases against the dictionary until there were some matches. If all sub-phrases matching failed, we then sought out exact matches for the component words in the phrase. For example, if a three-word phrase $w_1w_2w_3$ was missing in the dictionary, we searched the sub-phrases w_1w_2 and w_3; and if there was no match for w_1w_2, we searched w_1 and w_2w_3 in the dictionary. If none of the sub-phrases were found in the dictionary, we translated this phrase word-by-word by looking up each component word in the dictionary, and took the Chinese translations of all the component words in the phrase as the translation of the phrase. We submitted one run named Brkly-ECIR-LO-01 on this subtask. All content fields in the English topics were indexed. The Chinese translations were segmented into overlapping bigrams. The overall average precision was .2195 for the relaxed relevance judgment and .1908 for the rigid relevance judgment. The relatively poor performance can be mainly attributed to the limited coverage of the bilingual dictionary used for query translation. Many of the query terms were not translated because they are missing in the bilingual dictionary.

4 Japanese & English IR

The Japanese & English IR task is made up of six subtasks, covering monolingual, cross-language, and multilingual text retrieval. The test collection contains about 736,000 documents which are either abstracts of conference papers or extended summaries of Grant-in-Aid research report. About half of the abstracts written by the authors of the conference papers have English translations also provided by the authors. And about a quarter of the extended summaries of research reports have English translations. A typical document contains *title, author, abstract, keyword,* name of the conference fields. The keywords and their English translations are provided by the authors of the papers. A set of 49 topics in Japanese and in English are provided for these subtasks. A topic has a *title*, *description, narrative,* and *concept* fields. We will describe our participation in each of the six subtasks in the following subsections.

4.1 J-J Subtask

The J-J subtask concerns Japanese monolingual retrieval where the provided Japanese topics are searched against the Japanese documents only. Like in Chinese, word boundaries in Japanese are not marked. Thus, breaking sentences into words or other indexing units is needed in indexing. Our focus in this subtask was to compare the effectiveness of different segmentation techniques. The availability of pre-segmented copy of the test collection enables us to do comparison between bigram indexing and word-based indexing. We submitted five official runs for the J-J subtask named as Brkly1, Brkly2, Brkly3, Brkly4, and Brkly16, respectively. These five runs represent four different segmentation techniques. For the first two runs, Brkly1 and Brkly2, overlapping bigram indexing was applied to the topics and documents. The Brkly3 run used Chasen morphological analyzer to segment both the topics and the documents into words. The Brkly4 applied a non-overlapping bigram and unigram indexing method to segment the text. And the Brkly16 run used the pre-segmented topics and documents. Only words in the pre-segmented collection and topics were included in indexing. The average precision values for the five monolingual runs with respect to the first relevance judgment set are shown in column five in table 2. The table also shows what fields were indexed and what segmentation method was used for each run. All of our runs were produced without relevance feedback. At the first NTCIR workshop, we did not find the hiragara characters helpful in Japanese monolingual retrieval when overlapping bigram indexing is applied [3]. For Brkly1 run, we discarded any hiragara characters before generating bigrams.

Although bigram indexing is effective, the disadvantage is that the bigram index files are usually much larger than the word indexes generated from the same collection. We present an indexing method that breaks text into unigrams and bigrams that do not overlap. When the word length is limited to one or two characters, the number of possible ways to segment a sentence of n characters is given by the recurrence relation $N(n) = N(n-1)+N(n-2)$, where $N(n)$ is the number of ways to break a sentence of n characters into one or two-character words and $N(0) = 0, N(1) = 1, N(2) = 2$. For example, a sentence consisting of three letters can be split into non-overlapping unigrams and bigrams in one of three ways: 1) $S = C_1/C_2/C_3$, 2) $S = C_1C_2/C_3$, and 3) $S = C_1/C_2C_3$. For a segmented sentence $S = w_1w_2 \ldots w_m$, the probability of the sentence is computed as follows:

$$P(S) = P(w_1w_2 \ldots w_m) \quad (2)$$

$$= P(w_1)P(w_2) \ldots P(w_m) = \prod_{i=1}^{m} P(w_i) \quad (3)$$

where w_i is either a unigram or a bigram. Here we assume the words occur independently. We compute the probability of a sentence for all possible segmentations, then we choose the segmentation of the highest probability. The probability of a unigram or bigram is the maximum likelihood estimate in the test document. The probability of a one-character word (i.e., unigram) is estimated by $P(C_i) = \frac{N(C_i)}{N}$, and the probability of a two-character word (i.e., bigram) is estimated by $P(C_iC_j) = \frac{N(C_iC_j)}{N}$, where $N(C_i)$ is the number of times that character C_i occur in the corpus, $N(C_iC_j)$ is the number of times that string C_iC_j occurs in the corpus and N is the total number of times that all single character terms and all two-character terms occur in the corpus. When a sentence is short, one can easily enumerate all possible ways of segmenting the sentence and compute their associated probabilities, then choose the segmentation of the highest probability. But when a sentence is long, the number of possible segmentations is exponential, it is no longer practical to enumerate all possible ways of breaking the sentence and estimate their probabilities. However one can apply a dynamic programming technique to find out the most likely segmentation efficiently without computing the probabilities of all possible segmentations of a sentence. The best way of breaking a sentence of n characters can be recursively expressed as follows:

$$P(S_{1,n}) = MAX(P(S_{1,n-1})P(C_n), P(S_{1,n-2})P(C_{n-1}C_n))$$

where $S_{1,n} = C_1C_2 \ldots C_n$ and $P(S_{1,n})$ is the maximum probability of segmenting a sentence of n characters into one or two-character words. For Brkly4, we segmented the topics and documents into unigrams and bigrams that do not overlap with each other. The Chasen morphological analyzer was used to break the topics and documents into single words for Brkly3. Brly16 used the pre-segmented topics and documents, but only the words were included in indexing.

Table 3 shows the pairwise correlation values for four of the J-J runs that used the *title,description, narrative*, and *concept* fields in the topics. The pair of Brkly3 and Brkly16 has the highest correlation partly because both runs were performed on a word-based index. For Brkly3, the Chasen morphological analyzer was used to segment the documents and topics while the pre-segmented data was used in Brkly16

Run ID	Topic Fields Indexed	Document Fields Indexed	Topic/Document Segmentation Method	Average Precision
Brkly1	T,D,N,C	T,A,K,P	overlapping bigram	0.3602
Brkly2	D	T,A,K,P	overlapping bigram	0.2610
Brkly3	T,D,N,C	T,A,K,P	Chasen analyzer	0.3372
Brkly4	T,D,N,C	T,A,K,P	non-overlapping unigram and bigrams	0.3287
Brkly16	T,D,N,C	T,A,K,P	pre-segmented	0.3377

Table 2. This table shows the fields indexed in topics and documents and the segmentation methods used to break documents and topics into words.

run. As we expected, the Brkly1 run was more closely correlated with Brkly4 than with Brkly3 and Brkly16 because the indexing terms are overlapping bigrams for Brkly1 and are non-overlapping unigrams and bigrams for Brkly4. Overall when the average performance over all automatic runs on this subtask from all participating groups was poor, the performance of our runs was also poor. Our J-J monolingual runs achieved very low precision for topics 130, 132, 140, and 149. The performance on these four topics was also low on the average across all J-J automatic runs from all groups. The overlapping bigram indexing outperformed all other segmentation methods in our experiments. It was much more effective than word-based indexing on topics 105, 107, 139, and 146. For example, the precision on topic 146 is .6386 for Brkly1 which used bigram indexing, while the precision values for the other three runs are .2637, 0.2990, and 0.2764, respectively on the same topic.

	Brkly1	Brkly3	Brkly4	Brkly16
Brkly1	1.0000	0.8205	0.9146	0.8082
Brkly3	0.8205	1.0000	0.9116	0.9391
Brkly4	0.9146	0.9116	1.0000	0.9031
Brkly16	0.8082	0.9391	0.9031	1.0000

Table 3. Pearson Correlation Coefficients, N = 49, for all J-J runs.

4.2 E-E Subtask

We submitted two official runs, named Brkly5 and Brkly6, on this English monolingual subtask. Only the *description* field in the English topics was indexed for Brkly5, while the *title*, *description*, *narrative*, and *concept* fields were indexed for Brkly6. The English *TITE*, *ABSE*, *KYWE*, and *PJNE* fields in the English documents were indexed. In the preprocessing step, the stop words were removed from indexing in both topics and documents, the remaining words were changed to lower case and were reduced to their base form. We used an English morphological analyzer [7] to change the nouns in plural form into singular form, verbs into their infinitive form, and adjectives to their positive form. Again the same retrieval formula was applied without relevance feedback. The overall average precision is .3542 for Brkly5 and .2624 for Brkly6 with respect to the first set of relevance judgments.

4.3 J-E Subtask

Two official runs, named Brkly7 and Brkly8 respectively, were submitted on this Japanese to English cross-language retrieval subtask. The bilingual lexicon used in these two runs was created from the NTCIR-1 document collection. Our approach to Japanese-English cross-language retrieval is to create a bilingual lexicon from the documents with both Japanese and English keywords, then mapping each Japanese query term to its English equivalent. The English translations of all the query terms in a Japanese query are searched against the English collection. Thus we created two bilingual lexicons from the NTCIR-1 document collection. The first one called *bilingual keyword lexicon* was created from the keywords field only, while the second one called *bilingual word lexicon* was created from the *title* and *abstract* fields in the documents in NTCIR-1 collection.

4.3.1 Bilingual Keyword Lexicon

The existence of both Japanese and English keywords in the NTCIR-1 document collection enables us to build a bilingual lexicon. Most of the documents in the NTCIR-1 collection have both Japanese and English keywords assigned by the authors of the papers. The Japanese keywords in the *KYWD* field and the English keywords in the *KYWE* field are separated by two slash characters, making it easy to extract them.

Our bilingual lexicon was constructed from the Japanese and English keyword fields (i.e., the *KYWE* and *KYWD* fields) in the NTCIR-1 collection by pairing the Japanese keywords with the English keywords in the order they occur in the documents. That is, the first Japanese keyword is paired with the first English keyword in the same document, and the second Japanese keyword is paired with the second English keyword in the same document, and so on. This pair-

ing process terminates when either one of the keyword fields (KYWD and KYWE) is exhausted.

All of the Japanese/English keyword pairs are collected from the NTCIR-1 collection. The resulting bilingual lexicon consists of all the unique Japanese/English keyword pairs, each pair being associated with the number of occurrences in the NTCIR-1 collection.

When we paired the Japanese keywords with the English keywords in the same document, we were aware of the problems that the translations of Japanese keywords may not be consistent and complete, that the English translations and the original Japanese keywords in the same document may not be aligned properly and that the form of the English translations may not be normalized. For example, the words in the same English keyword is connected with hyphen in some cases, but not in other cases. Some of the Japanese keywords have more than one English translations because of inconsistency in translation of the the same terminology and misspellings in English.

4.3.2 Bilingual Word Lexicon

The second bilingual dictionary we created from the NTCIR-1 collection is based on words co-occurrence in the collection by first aligning the text at the sentence level. For each document with English translation, the abstracts in both English and Japanese are split into sentences. It is simple to break Japanese abstracts into sentences since there is a unique punctuation mark for sentence ending in Japanese, while it is more difficult to break English abstracts into sentences. Since the Japanese title and English title are marked in the source documents, each title in Japanese and its English translation are paired. For the abstracts, we used Gale and Church's program [10] to align Japanese sentences with English sentences. Gale and Church's method is based on a simple statistical model of sentence length measured in terms of characters. This method uses the fact that long sentences in a source language tend to be translated into long sentences in the target language and short sentences tend to be translated into short sentences. Even though it was originally developed to align English and French sentences based on sentence length in characters, we used the program without changing the model parameters to align Japanese and English sentences. The main differences are that we measured the sentence lengths in terms of bytes and that the blank spaces between English words within a sentence are also counted. In the test collection, a Japanese character is represented in two bytes, whereas an English character is represented in one byte. We generated about 959,000 pairs of Japanese and English sentences by aligning the sentences in title and abstract fields in the documents with English translations.

The Japanese sentences are segmented into words using a Japanese wordlist. If the English word 'E' is a translation of the Japanese word 'J', then one would expect that when the Japanese word 'J' is present in a Japanese sentence, its English translation 'E' would also appear in the paired English sentence. A number of statical measures, such as mutual information, likelihood ratio test-based [9, 8], have been developed to compute the association significance between two events. We used the likelihood ratio test-based measure developed by Dunning [8] to compute the association strength between a pair of Japanese/English words. From the aligned sentences, we constructed a contingency table for every pair of Japanese/English words as shown in table 4. where

	Japanese word	
English word	a	b
	c	d

Table 4. A contingency table for a pair of words.

a is the number of aligned sentences containing the pair of Japanese/English words; b is the number of aligned sentences containing the English word, but not the Japanese word; c is the number of aligned sentences containing the Japanese word, but not the English word; and d is the number of aligned sentences containing none of the word pair.

The association score between a Japanese word 'J' and an English word 'E' is computed as follows [8]

$$W(J,E) = 2[logL(p_1,a,a+b) + logL(p_2,c,c+d) - logL(p,a,a+b) - logL(p,c,c+d)]$$

where

$$logL(p,n,k) = klog(p) + (n-k)log(1-p)$$

and $p_1 = \frac{a}{a+b}$, $p_2 = \frac{c}{c+d}$, and $p = \frac{a+c}{a+b+c+d}$.

A total of 310,794 Japanese indexing terms and 278,800 English indexing terms (words) were generated from the parallel corpus of some 187,000 documents in the NTCIR-1 collection.

4.3.3 Query Translation

A simple method of translating queries into the target language is looking up each source language query word in a bilingual dictionary when such a dictionary is available. The translations for all source language query words can be combined to form the query to submit to the document collection in target language. In general such resources are not readily available,

and even if a general bilingual dictionary is available, its coverage on domain-specific terminological terms may be very limited. An alternative method of finding translation equivalents is to create a bilingual lexicon from parallel corpora when they are available. Then the bilingual lexicon can be used to look up source language query terms.

Our method of translating Japanese queries into English was to utilize the bilingual lexicons we had created from the NTCIR-1 collection. In translating Japanese queries into English, we first segment the queries into words using the dictionary-based longest-matching technique. Then for each Japanese word, the most frequent English translation is retained as the translation.

For Brkly7 and Brkly8 runs, we used only the bilingual keyword lexicon for query translation. The Japanese topics were segmented into words first using longest-matching method. The individual words were then looked up in the bilingual keyword lexicon. The English translation with the highest occurrence frequency was chosen as the translation of a Japanese word. The overall average precision is .2640 for Brkly7 and .2129 for Brkly8 using the first set of relevance judgment.

4.4 E-J Subtask

The English topics were processed in the same way as the English topics in English-to-Chinese subtask. First, the English topics were tagged with part-of-speech. Second, simple noun phrases were extracted based on predefined patterns of part-of-speech tags. We submitted three runs, named Brkly9, Brkly10, and Brkly11, respectively, on this English-to-Japanese cross-language retrieval. For Brkly10 run, only the *description* field was used. The other two runs indexed the *title*, *description*, *narrative*, and *concept* fields. For Brkly9, the same Japanese/English *bilingual keyword lexicon* created from the keyword fields in the NTCIR-1 collection was used to translate English phrases and words in the query into Japanese. For Brkly11, the missing words in the *bilingual keyword lexicon* were looked up in the *bilingual word lexicon*. The last column in table 5 shows the average precision for the three English-to-Japanese retrieval runs. The overall precision was improved moderately as a result of better coverage when both the keyword lexicon and the word lexicon were used in query translation for the Brkly11 run.

4.5 J-JE Subtask

The J-JE subtask involves searching Japanese topics against a collection comprised of mixed Japanese and English documents. The English documents are translations of the Japanese documents and the number of documents in Japanese and in English is about two to one. In general, one could approach this problem in one of two ways: combining monolingual retrieval results or combining queries in all document languages. The first step in both approaches is to translate the queries from the source language to all document languages. In this subtask, one needs only to translate the Japanese topics into English. With the first approach, a monolingual run is performed for each document language. Then the monolingual retrieval results are combined to produce a final ranked list of documents. With the second approach, the queries in all document languages are combined first. Then the pooled queries are searched against the mixed document collection. We took the first approach to the J-JE subtask. That is, we first translated the Japanese topics into English. We then performed one Japanese monolingual retrieval run using the subset of the Japanese documents which was the Brkly3 run; and one English monolingual run using the subset of the English documents which was the Brkly7 run. And finally we combined these two monolingual retrieval results to produce the final ranked list of documents. Our combining strategy is rather simple and is purely rank-based. It is based on the observations that the English documents are translations of the Japanese documents and the number of documents in Japanese and in English is approximately two to one. One would expect that the number of relevant Japanese documents should be also approximately twice the number of relevant English documents for a topic. The final ranked list of documents for Brkly12 run was produced by merging the monolingual Japanese retrieval result and the monolingual English retrieval result in two to one as illustrated in table 6. We submitted two official runs, named

Rank	Brkly3	Brkly7	Brkly12
1	J1	E1	J1
2	J2	E2	J2
3	J3	E3	E1
4	J4	E4	J3
5	J5	E5	J4
6	J6	E6	E2
...	...	...	...

Table 6. Document pooling strategy.

Brkly12 and Brkly13, on this subtask. The Brkly12 was produced by merging the Japanese monolingual run Brkly3 and the English monolingual run Brkly7. The final ranked list of documents for Brkly13 run was generated in the same way as for Brkly12 run by merging the Japanese monolingual run Brkly2 and the English monolingual run Brkly8. The average precision values for Brkly12 and Brkly13 are .2915 and .2211, respectively, with respect to the first set of relevance judgment.

Run ID	Topic Fields Indexed	Document Fields Indexed	Topic Segmentation Method	Topic Translation Method	Average Precision
Brkly9	T,D,N,C	T,A,K,P	Longest matching	keyword dictionary	0.2389
Brkly10	D	T,A,K,P	Longest matching	keyword dictionary	0.1489
Brkly11	T,D,N,C	T,A,K,P	Longest matching	keyword and word dictionaries	0.2544

Table 5. This table shows the fields indexed in topics and documents, the word segmentation methods for topics, and the topic translation method.

4.6 E-JE Subtask

The E-JE subtask is similar to the J-JE subtask except that the topics are in English. We took the same approach to this subtask. First, we translated the English topics into Japanese. Second, one English monolingual retrieval run and one Japanese monolingual retrieval run were performed. Finally, the monolingual retrieval results were merged into one final ranked list of documents. A small change in the document pooling strategy is that we put the English document before the Japanese documents each time when take one English document and two Japanese documents from the monolingual retrieval results as illustrated in table 7. The ratio of the number of documents in Japanese and in English is kept two to one. Two official runs, named

Rank	Brkly5	Brkly9	Brkly14
1	J1	E1	E1
2	J2	E2	J1
3	J3	E3	J2
4	J4	E4	E2
5	J5	E5	J3
6	J6	E6	J4
...	...	...	...

Table 7. Document pooling strategy.

Brkly14 and Brkly15, were submitted on this subtask. The Brkly14 was the result of combining the English monolingual run Brkly5 and the Japanese monolingual run Brkly9. And Brkly15 was generated by merging the English monolingual run Brkly6 and the Japanese monolingual run Brkly10. The average precision values with respect to the first set of relevance judgment are .2455 and .1570, respectively.

5 Concluding Remarks

The overall precision on the long and short queries in the Chinese monolingual subtask demonstrated the effectiveness of the bigram indexing in Chinese information retrieval. The results on Japanese monolingual retrieval show that bigram indexing outperformed word indexing. The performance of our cross-language retrieval runs was limited by the many missing translations of the query words. The use of the bilingual word dictionary in addition to the keyword dictionary in query translation resulted in a moderate increase in overall precision on the English to Japanese cross-language retrieval subtask. On the multilingual retrieval tasks, we presented a rank-based document pooling strategy for merging monolingual retrieval results to produce a single ranked list of documents in Japanese and English. The results on retrieval from mixed document collection shows this simple document combining strategy worked reasonably well.

6 Acknowledgements

This research was supported by DARPA under research grant N66001-00-1-8911 as part of the DARPA Translingual Information Detection, Extraction, and Summarization Program (TIDES).

References

[1] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.

[2] A. Chen. Phrasal translation for english-chinese cross language information retrieval. In *Workshop on English-Chinese Cross Language Information Retrieval at the 2000 International Conference on Chinese Language Computing*, pages 195–202, Chicago, Illinois, USA, July 2000.

[3] A. Chen, F. Gey, K. Kishida, H. Jiang, and Q. Liang. Comparing multiple methods for japanese and japanese-english retrieval. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 49–58, Tokyo, Japan, 1999.

[4] A. Chen, J. He, L. Xu, F. C. Gey, and J. Meggs. Chinese text retrieval without using a dictionary. In *20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, USA.*, 1997.

[5] A. Chen, K. Kishida, H. Jiang, Q. Liang, and F. C. Gey. Automatic construction of a japanese-english lexicon and its application in cross-language information retrieval. In *Joint ACM DL/ACM SIGIR Workshop on Multilingual Information Discovery and AccesS (MIDAS)*, Berkeley, California, USA, Aug. 1999.

[6] W. S. Cooper, A. Chen, and F. C. Gey. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 57–66, March 1994.

[7] M. Z. Daniel Karp, Yves Schabes and D. Egedi. A freely available wide coverage morphological analyzer for english. In *Proceedings of COLING*, 1992.

[8] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19:61–74, March 1993.

[9] W. A. Gale and K. W. Church. Identifying word correspondences in parallel texts. In *Proceedings of the Fourth DARPA Speech and Natual Language Workshop*, pages 152–157, Pacific Grove, CA, 1991.

[10] W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19:75–102, March 1993.

[11] F. C. Gey and A. Chen. Phrase discovery for english and cross-language retrieval at trec-6. In D. K. Harman, editor, *Text Retrieval Conference (TREC-6)*, pages 637–648, 1997.

[12] F. C. Gey, A. Chen, J. He, L. Xu, and J. Meggs. Term importance, boolean conjunct training, negative terms, and foreign language retrieval: Probabilistic algorithms at trec-5. In D. K. Harman, editor, *Text Retrieval Conference (TREC-5)*, 1996.