Comparing multiple methods for Japanese and Japanese-English text retrieval

Aitao Chen*, Fredric C. Gey†, Kazuaki Kishida‡, Hailing Jiang*and Qun Liang*

*School of Information Management and Systems
102 South Hall
University of California at Berkeley, CA 94720, USA
{aitao,hjiang1,qliang}@sims.berkeley.edu

†UC Data Archive & Technical Assistance (UC DATA)
University of California at Berkeley, CA 94720, USA
gey@ucdata.berkeley.edu

‡Faculty of Cultural Information Resources, Surugadai University,
698 Azu, Hanno, 357-8555 Japan
kishida@surugadai.ac.jp

Abstract

The NACSIS collection of Japanese scientific documents (with English titles) provides a solid foundation for information retrieval research into 1) segmentation methods for Japanese text, 2) effective methods for monolingual Japanese retrieval, and 3) Japanese-English cross-language retrieval. This paper compares multiple methods for Japanese and Japanese-English text retrieval. Our focus is on accurate methods to segment Japanese test and on the construction of a large bi-lingual Japanese-English lexicon. In cross-language retrieval we have used these methods to compare a targeted dictionary-based approach to CLIR against a machine translation approach.

In monolingual retrieval we have found that overlapping bigrams for both query and document perform better than dictionary lookup where the dictionary is abridged. Our NTCIR cross-language results show that creation of a bi-lingual lexicon tailors the retrieval to particular domain, and can improve average precision by fifty percent over general machine translation, which does not benefit from specialized domain knowledge provided by lexicon construction from a parallel corpus.

1 Introduction

This paper compares multiple methods for Japanese and Japanese-English retrieval. We participated in the ad hoc and cross-lingual tasks in which we tested two word segmentation methods and two query translation methods. This work builds on our earlier work [3, 1, 7] on full-text monolingual and cross-language information retrieval undertaken through our participation in the Text REtrieval Conferences (TREC)[1].

2 Test Collection

The data collection we used in all of our experiments reported here is the *NACSIS Test Collection 1* [10](NTCIR-1) of some 330,000 documents, 50 queries and their relevance judgments. The test collections has three parts: *ntc1-je0*, *ntc1-j0*, and *ntc1-e0*. About 187,000 of the documents in the ntc1-je0 collection contain English translations. The ntc1-j0 collection consists of documents in the ntc1-je0 collection without the English fields, and the ntc1-e0 collection consists of the documents in the ntc1-je0 collection without the Japanese fields. The documents are summaries of papers presented at conferences hosted by Japanese academic societies. The collection covers a variety of topics, such as chemistry, electrical engineering, computer science, linguistics, library science, and so on. 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 topic has a title, description, narrative, and concept fields. Some of the topics also contain concept terms and acronyms in English.

This test collection is unique because many of the documents have author assigned keywords and their English translations. Figure 1 shows a sample topic and figure 2 shows a sample document.

[1] http://trec.nist.gov/

<TOPIC q=0043>

<TITLE>
動画像圧縮センサ
</TITLE>

<DESCRIPTION>
動画像圧縮を行なう知能化イメージセンサに関する研究が知りたい。
</DESCRIPTION>

<NARRATIVE>
画像を扱うシステムにおいて、高精細、高フレームレート等の高レート化に伴い、読み出し、転送時における遅延が問題となってきた。既存のシステムにおいては画像獲得と画像処理はほぼ完全に分離されているのに対し、イメージセンサ上での画像圧縮機能を実現し、画像取得と画像処理をより密接に関連させてこれらのボトルネックを解消しようというアプローチが検討され始めている。動画像の圧縮をイメージセンサ上で行なうことを目的とした論文が欲しい。画像処理をしているが圧縮はしていないものは不正解。蓄積されている動画像に対して圧縮処理をするものは要求を満たさない。研究動向調査のため。
</NARRATIVE>

<CONCEPT>
<J.CONCEPT>
a. コンピュテーショナルセンサ, 知能化センサ, インテリジェントセンサ,
b. 画像センサ,
c. 動画像圧縮
</J.CONCEPT>
<E.CONCEPT>
a. Computational sensor, smart sensor, Intelligent sensor,
b. Image Sensors,
c. Video compression, image compression

Figure 1: A sample topic.

2.1 Document Ranking

The document ranking formula we used in all of our retrieval runs was Berkeley's TREC-2 formula [3]. 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 [6]. The logodds of relevance of document D to query Q is given by

$$\log O(R|D,Q) = log\frac{P(R|D,Q)}{P(\overline{R}|D,Q)} \quad (1)$$

$$= -3.51 + \frac{1}{\sqrt{N}+1}\Phi + .0929 * N \quad (2)$$

$$\Phi = 37.4\sum_{i=1}^{N}\frac{qtf_i}{ql+35} + 0.330\sum_{i=1}^{N}\log\frac{dtf_i}{dl+80} - 0.1937\sum_{i=1}^{N}\log\frac{ctf_i}{cl} \quad (3)$$

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 variables in the document ranking formula are defined in table 1. The summation in equation (3) is carried out over the matching terms between the document and the query. The relevance probability of document D with respect to query Q can be written as follows given the logodds of relevance.

$$P(R|D,Q) = \frac{1}{1+e^{-logO(R|D,Q)}} \quad (4)$$

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 training data to the logistic regression model using a statistical software package.

3 Ad hoc/Monolingual Tasks

In most information retrieval systems, the documents and queries are represented in words. To represent

<REC>
<ACCN>gakkai-0000185278</ACCN>
<TITL TYPE="kanji">動画像圧縮イメージセンサの検討</TITL>
<TITE TYPE="alpha">On-sensor Video Compression</TITE>
<AUPK TYPE="kanji">大野 洋 / 浜本 隆之 / 相澤 清晴 / 羽鳥 光俊 / 山崎 順一 / 丸山 裕孝</AUPK>
<AUPE TYPE="alpha">Ohno,Hiroshi / Hamamoto,Takayuki / Aizawa,Kiyoharu / Hatori,Mitsutoshi / Yamazaki,Jun-ichi / Maruyama,Hirotaka</AUPE>
<CONF TYPE="kanji">画像応用研究会</CONF>
<CNFE TYPE="alpha">Technical Group on Applied Image Processing and System</CNFE>
<CNFD>1994. 08. 26</CNFD>
<ABST TYPE="kanji"><ABST.P>画像を扱う既存のシステムにおいては,画像獲得と画像処理はほぼ完全に分離している。ところが画像技術の応用分野が広がるにつれ,イメージセンサに対して,高レート化,高機能化が要求されるようになってきた。これらの要求に従来の枠組で対応していくと,画像情報を1次元の時系列信号として転送する場合,転送遅延がボトルネックとなってしまう。この問題に対して,センサ上で一部(あるいは全て)の処理を実行し,画像取得と画像処理をより密接に関連させて解決しようというアプローチが検討され始めている。</ABST.P><ABST.P>我々はセンサ上で適切な画像圧縮を施すことで,取得画像の高レート化(高速化,高精細化)に対応することを考えている。本稿では,センサ上での動画像圧縮のためのアルゴリズム,およびそのチップへの実装について論じる。</ABST.P></ABST>
<ABSE TYPE="alpha"><ABSE.P>In this paper,we propose new computational image sensors which compress image signal in the process of image acquisition.Conditional replenishment is used to reduce the band-width necessary for image read-out.We also describe about the design of the experimental chip.This chip has an extensible,parallel,architecture.</ABSE.P></ABSE>
<KYWD TYPE="kanji">画像センサ // コンピュテーショナルセンサ // 画像圧縮 // 画像符号化</KYWD>
<KYWE TYPE="alpha">Image Sensors // Computational Sensors // Image Compression // Image Coding</KYWE>
<SOCN TYPE="kanji">テレビジョン学会</SOCN>

Figure 2: A sample document.

Japanese documents in words, the documents need to be segmented into words since the word boundaries are not marked in Japanese text. Our focus in ad hoc task was comparing the retrieval performances of different word segmentation techniques.

We submitted three official runs for the ad hoc task: BKJJBIFU, BKJJBIDS, and BKJJDCFU. The document collection we used was the ntc1-j0 collection. After submitting the official results for the ad hoc task, we realized that we should have used the ntc1-je0 collection. Since the ntc1-j0 collection is the same as the ntc1-je0 collection when the English fields are ignored, we would have produced the same results had we indexed the ntc1-je0 collection with the English text in the documents ignored. Because we used ntc1-j0 collection for the monolingual task, the results from the same three runs were also submitted under the monolingual task category.

3.1 Indexing Dictionary

Since the word boundaries in Japanese writing are not marked, segmenting Japanese text into words usually comes as the first step in indexing. One of the word segmentation methods is the *dictionary-based longest matching* which matches the initial string of characters against the dictionary entries and takes the initial string that matches the longest entry in the dictionary as a word. In general, achieving high accuracy in word segmentation will require a dictionary of wide coverage over the text to segment. As mentioned above, the NTCIR-1 collection consists of the summaries of technical papers where technical terms are prevalent. The richness in technical terms in the text poses a problem to word segmentation since the technical terms are often missing in a general language dictionary of reasonably large size. The automatic extraction of terms from the text to segment could play an important role in building a dictionary for word segmentation.

We created a dictionary (perhaps term list would be more appropriate) by 1) merging the words in the dictionary in the Chasen morphological analyzer[11], the Japanese words in the *edict* dictionary, and Japanese terms extracted from the Japanese keyword field (i.e. KYWD field) in the documents of the ntc1-j0 collection, and 2) stripping all hiragana characters from the entries in the combined word list. The extracted terms from the Japanese keyword field are kanji and katakana fragments. In this paper, a Japanese term may refer to the root of a word, part of a word, a word, a compound, and a phrase. Our Japanese dictionary has 419,741 entries, consisting of the kanji fragments or the

N	is the number of terms common to both query and document,
qtf_i	is the occurrence frequency within a query of the ith matching term,
dtf_i	is the occurrence frequency within a document 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.

Table 1: Definitions of the variables in the document ranking formula.

Run ID	Topic Fields Indexed	Document Fields Indexed	Category	Topic/Document Segmentation Method	Document Collection
BKJJBIFU	TITLE, DESCRIPTION, NARRATIVE, J.CONCEPT, A.CONCEPT	TITL, ABST, KYWD	Automatic	Bigram	ntc1-j0
BKJJBIDS	DESCRIPTION	TITL, ABST, KYWD	Automatic	Bigram	ntc1-j0
BKJJDCFU	TITLE, DESCRIPTION, NARRATIVE, J.CONCEPT, A.CONCEPT	TITL, ABST KYWD	Automatic	Longest-matching	ntc1-j0

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

katakana fragments. Most of the dictionary entries were extracted from the Japanese keyword field. This dictionary was used to segment documents and topics in the retrieval runs in which the longest-matching algorithm was used to break chunks of kanji and katakana characters into words. It was also used in the Japanese-English cross language retrieval to segment topics before the Japanese query words were translated into English.

3.2 Topic and Document Indexing

Four sets of characters are used in Japanese writing: kanji, katakana, hiragana, and Roman characters. The characters are mixed in writing. Like in Chinese, word boundaries in Japanese writing are not marked. The hiragana characters are not content-bearing terms in most cases, thus they were excluded from indexing, resulting in fragments consisting of only either kanji characters or katakana characters.

Table 2 presents the fields in documents and topics that were indexed for each retrieval run. All the English words mixed in the Japanese text were retained in lower case. The English words in the English concept field (E.CONCEPT) in the topics were not indexed. Only the *TITL*, *ABST*, and *KYWD* fields in the ntc1-j0 collection were indexed. The text in the *TITL*, *ABST*, and *KYWD* fields were split into fragments of text consisting of kanji and katakana characters only. Everything else including hiragana characters was stripped in the first step of indexing. The kanji and katakana fragments were further segmented into smaller indexing terms. For the retrieval runs 'BKJJBIFU' and 'BKJJBIDS', the kanji and katakana fragments were further segmented into overlapping bigrams, and for the retrieval run 'BKJJDCFU', they were segmented into indexing terms by using the maximum-matching (also called longest-matching) method [1] against our Japanese dictionary.

3.3 Results

Table 3 presents the precision values at 11 recall points, the average precision values, and the number of relevant documents retrieved for the BKJJBIFU, BKJJBIDS, and BJKKDCFU runs, which were all automatic. The average for each run was taken over 50 test topics.

Table 4 shows the precision values at 11 recall levels, the average precision over 39 test topics, and total number of relevant documents retrieved for the same

three runs. The partial relevance file for the monolingual retrieval task was used. . The results in table 4 show bigram segmentation has substantially outperformed the dictionary-based longest segmentation. Despite its simplicity, the bigram segmentation method combined with the logistic regression-derived ranking formula performed well on the NTCIR-1 collection.

The relative poor performance of the dictionary-based segmentation may be attributed to the poor quality of the dictionary used to segment text. We noticed in our dictionary that there are many long kanji and katakana fragments that should be broken into smaller components.

4 Cross-Lingual Task

Cross-language information retrieval usually is carried out by translating queries, or translating documents, or translating both the documents and queries to a third language [9, 12]. Queries can be translated by using machine translation systems or looking up bilingual dictionaries. The coverage of the bilingual dictionary used to translate queries could have large impact on the performance of a cross-language retrieval system. 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 the test collection itself or some parallel or comparable text corpora that is similar in content to the test collection. Then the bilingual lexicon can be used to look up source language query terms. Our approach to Japanese-English cross-language retrieval is creating 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 (ntc1-e0).

The existence of both Japanese and English keywords enables us to build a bilingual lexicon from the collection itself.

4.1 Bilingual Lexicon

Most of the documents in the ntc1-je0 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 ntc1-je0 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 pairing 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 ntc1-je0 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 ntc1-je0 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.

Figure 3 presents a small fragment of the bilingual lexicon (Japanese/English keyword pairs) derived from the ntc1-je0 collection. The first column is the number of times that a Japanese/English keyword pairs occurs in the collection. The second column is the Japanese/English pair separated by a vertical bar. As the fragment of the lexicon shows, the same Japanese keyword has several translations, such as *graphic compression*, *graphic data compression*, *image compression*, *image data compression*, *image/video compression*, *picture compression*, et al.

4.2 Query Translation

Our method of translating Japanese queries into English is looking up bilingual lexicon we created from the ntc1-je0 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. One of the problems in cross-language information retrieval (CLIR) is to decide how many translations to retain [8]. Since the test collection consists of summaries of technical papers, we assume that each Japanese indexing term, in general, has only one English translation, which may include more than one English word. The English translations were submitted to the ntc1-e0 collection.

Figure 4 shows the segmentation results of topic 43 using the longest-matching method and the transla-

Run ID	BKJJBIFU	BKJJBIDS	BKJJDCFU
Recall Level	Precision	Precision	Precision
at 0.00	0.8848	0.7751	0.8325
at 0.10	0.8020	0.5800	0.7228
at 0.20	0.7020	0.4623	0.5817
at 0.30	0.5882	0.3896	0.4992
at 0.40	0.5323	0.3207	0.4119
at 0.50	0.4557	0.2722	0.3405
at 0.60	0.3653	0.2150	0.2843
at 0.70	0.2625	0.1809	0.2093
at 0.80	0.1990	0.1372	0.1401
at 0.90	0.1219	0.0777	0.0630
at 1.00	0.0552	0.0541	0.0414
Average Precision	0.4350	0.2927	0.3536
Relevant Retrieved	1628	1226	1462

Table 3: Evaluation results for the ad hoc retrieval task. There are 2345 relevant documents for all 50 test queries in the partial relevant file.

Run ID	BKJJBIFU	BKJJBIDS	BKJJDCFU
Recall Level	Precision	Precision	Precision
at 0.00	0.8883	0.8053	0.8350
at 0.10	0.8253	0.6133	0.7092
at 0.20	0.7066	0.4575	0.5548
at 0.30	0.5961	0.3909	0.4636
at 0.40	0.5210	0.3144	0.3833
at 0.50	0.4497	0.2646	0.3163
at 0.60	0.3612	0.2108	0.2532
at 0.70	0.2802	0.1713	0.2001
at 0.80	0.2186	0.1292	0.1344
at 0.90	0.1180	0.0538	0.0520
at 1.00	0.0430	0.0268	0.0202
Average Precision	0.4378	0.2888	0.3329
Relevant Retrieved	1457	1226	1293

Table 4: Evaluation results for Japanese monolingual retrieval task. There are 2101 relevant documents for all 39 test queries in the partial relevant file.

tion results by bilingual dictionary lookup. The major portion of the dictionary used to segment the topic and the entire bilingual dictionary were derived from the NTCIR-1 collection. The English equivalent of a Japanese term is its most frequent translation. A Japanese term is not translated when it is missing in the bilingual lexicon.

The words in the ntc1-e0 collection were stemmed using the SMART [2] system stemmer and the stopword list included in the SMART system was used to remove non-content bearing words. The translated English query words were processed in the same way as the English documents.

[2]Available via ftp at ftp.cs.cornell.edu/pub/smart.

4.3 Results

We submitted five official runs in cross-lingual task, which were all automatic. Table 5 presents the evaluation results for all five runs. The average precision was computed over 39 test topics, and the partial relevance files were used in the evaluation. Table 6 shows what fields in the topics and documents in the *ntc1-e0* collection were indexed. For the runs BKJEBKFU, BKJEBDFU, and BKJEBDDS, the query terms were translated by looking up the bilingual lexicon that we created from the *ntc1-je0* collection.

The only difference between the two runs BKJEBKFU and BKJEBDFU is that the first run includes the English concept terms and the second does not. The run BKJEBDDS indexes only the description field in the topics. The topics in these three runs were translated into English using the same bilingual lexicon derived from the Japanese and English keyword fields.

Figure 3: A fragment of the Japanese/English keyword pairs created from the ntc1-je0 collection. The first column is the number of occurrences in the collection, and the second column is the Japanese/English keyword pairs, separated by a vertical bar.

The test topics in the BKJEMTFU run were translated into English using a machine translation system [3]. The cross-lingual run BKJEECFU, which was a mistake, used only the English concept in the topics as the queries submitted to the English test collection. For the BKJEECFU run, we intended to use the small *edict* Japanese-English dictionary to translate the queries into English to show what impact the dictionary coverage might have on the final retrieval performance in cross-lingual retrieval.

The method of aligning the keywords in Japanese and English in the order they occur in the documents and then choosing the English translation most frequently found in the collection for a Japanese keyword is simple and effective as our cross-language results presented in table 5 show. However, this method can be applied only when the documents containing keywords in both the source and target languages are available for creation of bilingual lexicon.

After we submitted the official runs for the cross-lingual task, we carried out additional experiments in which no English concept terms in the topics were retained in the queries and no keywords in Japanese and English were utilized to create the bilingual lexicon that was used to translate Japanese query terms into English. A large parallel test collection is hard to come by and it is even more difficult to have a large parallel test collection also with bilingual keywords.

For the additional run, we applied the sentence alignment technique developed by Gale and Church [5] to align the abstracts in Japanese and English on the sentence level. Then we used the measure of association between two events developed by Dunning [4] to find the most likely English translations of the Japanese terms. To translate a Japanese word into English, we computed the association strength between the Japanese word and every English words that co-occur with the Japanese word in at least one aligned sentences pair. The English words are ranked by their association scores with the Japanese word, and up to four top English words are taken as the translation of the Japanese word. The number of English words to retain is heuristically determined based on the type of Japanese word (i.e. kanji or katakana) and the number

[3] we are grateful to Kevin Knight and Ed Hovy at Information Science Institute in the University of Southern California for kindly translating the test queries into English using their machine translation system.

1 1 画像圧縮 image compression
2 1 不正
3 1 時 time
4 1 始
5 1 検討 study
6 1 研究 research
7 1 コンピューショナルセンサ computational sensor
8 1 圧縮処理 compression technique
9 2 高
10 2 行 line
11 1 レート rate
12 1 欲
13 1 要求 demand
14 1 問題 problem
15 1 目的 purpose
16 1 滴
17 1 密接
18 2 動画像 moving picture
19 2 システム system
20 1 ボトルネック bottleneck
21 1 フレームレート frame rate
22 1 関
23 1 機能 function
24 1 転送 propagation
25 2 圧縮 compression
26 1 インテリジェントセンサ intelligent sensor
27 1 実現 implementation
28 1 解
29 1 化
30 1 伴
31 1 解消
32 3 画像 image
33 3 動画像圧縮 video compression
34 1 遅延 delay
35 1 蓄積 storage
36 1 扱
37 1 分離 separation
38 1 読
39 3 イメージセンサ image sensor
40 1 アプローチ approach
41 1 等
42 1 論文 article
43 3 センサ sensor
44 1 知
45 2 対 tai
46 1 完全
47 1 関連 relationship
48 1 既存
49 2 知能化 intelligence
50 1 研究動向調査 survey of trends of research
51 1 獲得 acquisition
52 1 高精細 high definition
53 3 画像処理 image processing
54 1 画像取得 image acquisition

Figure 4: The segmentation and translation results of topic 43. The topic was segmented using the longest-matching method and the translation was bilingual dictionary lookup. The English translation of a Japanese term is its most frequent translation found in the NTCIR-1 collection. Each entry has four parts (in order): 1) sequence number, 2) frequency of a Japanese term in topic 43, 3) Japanese terms resulted from word segmentation, and 4) the English equivalents of the Japanese terms when the Japanese terms are found in the bilingual lexicon or empty when they are absent from the bilingual lexicon.

of characters in the Japanese word. More details on the creation of the bilingual lexicon from only the title and abstracts in Japanese and English and the translation of the topics are presented in [2]. The translated queries were submitted to the English collection (ntc1-e0) to retrieve 1000 documents for each query. The average precision over 39 test queries for this run was 0.3141.

5 Conclusions

For monolingual Japanese retrieval we have found, perhaps surprisingly, that simpler is better. Overlapping bigram segmentation of kanji and katakana text fragments outperformed dictionary (lexicon) based segmentation by more than 30%. This is due to both the incompleteness of the dictionary and its phrasal nature, i.e, we had no way to semantically decompose longer text sequences into meaningful words. For cross-language retrieval, however, phrasal segments provide greater precision of translation.

The retrieval performance of CLIR could be affected by a number of factors, such as the quality of the translation in parallel corpora, the accuracy of word segmentation in Japanese, the effectiveness of the document ranking formula, and so on. The incompleteness and inconsistency in translation and misspellings of English words could degrade the quality of the bilingual lexicon, which will eventually degrade the retrieval performance of CLIR.

6 Acknowledgements

We would like to thank Noriko Kando at NACSIS, Japan for making the NACSIS Test Collection 1 (NTCIR-1), test queries, and their relevance judgments available to us for research purpose. All authors are participants of the NTCIR Workshop (http://www.rd.nacsis.ac.jp/~ntcadm/workshop/work-en.html/). This research was supported by the Information and Data Management Program of the

Run ID	BKJEBKFU	BKJEBDFU	BKJEBDDS	BKJEMTFU	BKJEECFU
Recall Level	Precision	Precision	Precision	Precision	Precision
at 0.00	0.8630	0.7915	0.5466	0.5660	0.2632
at 0.10	0.6799	0.6353	0.4464	0.4708	0.2155
at 0.20	0.6159	0.5613	0.3806	0.3576	0.1776
at 0.30	0.5130	0.4839	0.3098	0.2503	0.1528
at 0.40	0.4379	0.4085	0.2612	0.2209	0.1286
at 0.50	0.3996	0.3640	0.2151	0.1621	0.1112
at 0.60	0.2918	0.2806	0.1776	0.1130	0.0814
at 0.70	0.2497	0.2161	0.1189	0.0858	0.0701
at 0.80	0.1818	0.1475	0.0812	0.0663	0.0515
at 0.90	0.0757	0.0677	0.0415	0.0441	0.0364
at 1.00	0.0527	0.0443	0.0250	0.0366	0.0162
Average Precision	0.3755	0.3438	0.2205	0.1925	0.1111
Relevant Retrieved	808	794	618	722	247

Table 5: Evaluation results of the Japanese cross-lingual retrieval runs. There are 1025 relevant documents for all 39 test queries in the partial relevant file.

National Science Foundation under grant IRI-9630765. It was also supported by DARPA (Department of Defense Advanced Research Projects Agency) under research contract N66001-97-C-8541, AO-F477.

References

[1] Aitao Chen, Jianzhang He, Liangjie Xu, Fredric C. Gey, and Jason 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.

[2] Aitao Chen, Kazuaki Kishida, Hailing Jiang, Qun Liang, and Fredric 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.

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

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

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

[6] 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.

[7] Fredric C. Gey and Aitao 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.

[8] Gregory Grefenstette, editor. *Cross-language information retrieval.* Kluwer Academic Publishers, Boston, MA, 1998.

[9] David A. Hull and Gregory Grefenstette. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[10] Noriko Kando and et al. NTCIR: NACSIS Test Collection Project. In *the 20th Annual BCS-IRSG Colloquium on Information Retrieval Research, March 25-27, 1998, Autrans, France*, 1998.

[11] Y. Matsumoto et al. Japanese morphological analyzer chasen 1.5. 1997.

[12] Douglas W. Oard. A comparative study of query and document translation for cross-language information retrieval. In *Machine Translation and the Information Soup. Third Conference of the Association for Machine Translation in the Americas*, pages 472–83, 1998.

Run ID	Topic Fields Indexed	Document Fields Indexed	Category	Topic Segmentation Method	Topic Translation Method	Document Collection
BKJEBKFU	TITLE, DESCRIPTION, NARRATIVE, J.CONCEPT, A.CONCEPT, E.CONCEPT	TITE, ABSE KYWE	Automatic	Longest-matching	Dictionary	ntc1-e0
BKJEBDFU	TITLE, DESCRIPTION, NARRATIVE, J.CONCEPT, A.CONCEPT	TITE, ABSE, KYWE	Automatic	Longest-matching	Dictionary	ntc1-e0
BKJEBDDS	DESCRIPTION	TITE, ABSE, KYWE	Automatic	Longest-matching	Dictionary	ntc1-e0
BKJEMTFU	TITLE, DESCRIPTION, NARRATIVE, J.CONCEPT, A.CONCEPT	TITE, ABSE, KYWE	Automatic	None	Machine Translation	ntc1-e0
BKJEECFU	E.CONCEPT	TITE, ABSE, KYWE	Automatic	None		ntc1-e0

Table 6: This table shows the fields indexed in topics and documents, the word segmentation methods for topics and queries, and the topic translation method.