Cross-Lingual Information Retrieval using Automatically Generated Multilingual Keyword Clusters

Noriko Kando　　Akiko Aizawa

Research and Development Department

National Center for Science Information Systems (NACSIS)

Tokyo, Japan

Email: {kando,akiko}@rd.nacsis.ac.jp

URL: http://www.rd.nacsis.ac.jp/~{kando,akiko}/

ABSTRACT

We propose an approach for cross-lingual information retrieval (CLIR) using the automatically generated multilingual keyword clusters based on graph theory, and show its effectiveness in IR and CLIR. The graph theoretic method has advantages that low-frequency keywords can be maintained for later use in IR. The results of retrieval against *NACSIS Test Collection 1* showed that query expansion using the clusters improved the search effectiveness in monolingual retrieval, by 13.2%, 14.2% at the level of “Relevant” and “Partially Relevant”, respectively. The search effectiveness of CLIR attained levels of 52.4% , 65.7% of the results for monolingual retrieval with "Relevant", and "Partially Relevant", respectively without any manual interaction during the retrieval. Future studies are also discussed.

Keywords

Cross-lingual information retrieval, English and Japanese, bilingual keywords clusters, graph theory, Japanese text retrieval, character-based indexing, word-, and phrase-based query segmentation

1. Need for Cross-Lingual Information Retrieval

In this paper we propose an approach for cross-lingual information retrieval (CLIR) using automatically generated multilingual keyword clusters based on graph theory, and show its effectiveness in CLIR and monolingual retrieval. We generate bilingual keyword clusters automatically, using the readily available bilingual domain-specific corpora with practically reasonable computational cost. More specifically, we use, as the corpora, Japanese and English keywords, which are assigned to scientific papers by the authors. The majority of scientific papers published in Japan have both Japanese and English keywords assigned by the authors. These bilingual keywords are readily available in databases for a great many subject domains and they are often more specific than the terms listed in the dictionaries.

The graph-theoretic method also has advantages of that the low-frequency keywords can be treated properly and maintained for later use in retrieval. As our main purpose of the keyword clustering is to use in retrieval, we generate clusters of related terms, which are expected to be effective in improving search performance, rather than only identifying the correct English-Japanese term pairs. A cluster may contain both English and Japanese terms, so that it is usable for both translating query terms into other languages in CLIR, and query expansion in monolingual retrieval. We also conducted automatic phrase identification and synonyms' operation in order to reduce the translation ambiguity.

The explosive growth of online documents, in particular due to the internet, has increased the need for information retrieval (IR) systems that cross language boundaries. CLIR systems offer the potential for users to submit a query in one language and retrieve documents from a collection in other languages. CLIR would be useful for people who are not fluent enough to create and submit a query in a foreign language, but who can read the language well enough to understand document contents and judge their relevance. CLIR also helps to reduce the cost of manual translation by discarding irrelevant documents before translation.

Moreover in Japanese scientific and technical documents, a concept can be expressed in four forms: English terms in original spelling (in roman alphabets), acronyms (in roman alphabets), transliterated form (in Katakanas, which are phonetic syllables in the Japanese language), and Japanese terms. CLIR technique is also expected to solve the problem of word mismatch caused by such variation in Japanese texts (Kando, 1997). A different approach has been reported for the similar problem of the English terms in the Korean texts (Myaeng & Kwon, 1997).

In the following, in section 2, we describe the method of the keyword clustering, used in this paper. In section 3, we present the experimental methods. Section 4 reports the results of CLIR and monolingual retrieval using the cluster. The search effectiveness is tested against the test version of the *NACSIS Test Collection 1* (Kando et al, 1998a), which contains about 330,000 documents. Section 5 discusses related works in CLIR and future studies.

2. Multilingual Keyword Clusters

In this section, we briefly describe our method for generating Japanese and English bilingual keyword clusters from the keyword lists assigned to academic papers by the authors. The detailed procedure has been reported in (Aizawa et al, 1998a, 1998b). Figure 1 shows the overall procedure.

The feature of our clustering strategy is that we adopt a graph-theoretic method instead of statistical ones, which currently

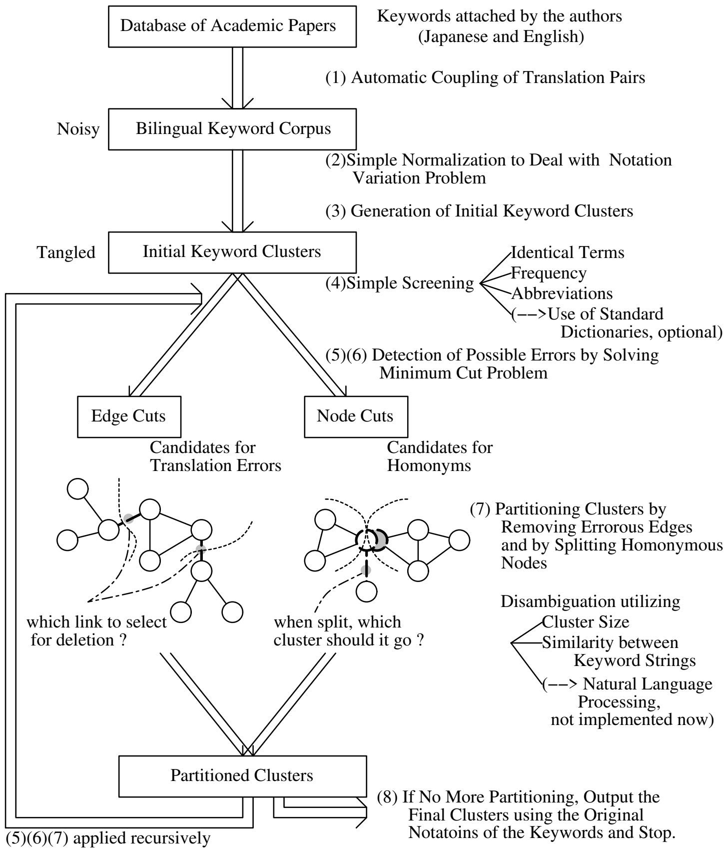


Figure 1: Overview of the proposed keyword clustering procedure

seem dominant in corpus-based approaches. The basic idea here is, the bilingual keyword pairs constitute a tangled graph of Japanese and English keywords; as such, the clustering problem can be regarded as a problem of partitioning the original keyword graph by eliminating wrongly generated links from the graph; this problem can be transformed into the *minimum cut problem* in the graph theory.

2.1 Extraction of Japanese-English Keyword Pairs

As the first step, we extracted Japanese-English Keyword pairs from the corpus. The basic data used in the current study are the Japanese and English keywords assigned by the authors to their papers, extracted from the *NACSIS Academic Conference Database* (NACSIS, 1997) in the *NACSIS Test Collection 1*, which we used in the retrieval experiments reported in this paper). We selected 28,122 (of 338,668 in the whole collection) papers in the field of computer science. Of the papers selected, 26,060 (about 93 %) have equal numbers of Japanese and English keywords. These keywords do not use any controlled vocabulary. An example is:

Japanese: 言語横断検索 / 用語クラスタ / グラフ理論

English: cross-lingual information retrieval / keyword clusters / graph theory

Though the authors of the papers are not explicitly required to provide precise conceptual correspondence between the two languages, our manual examination of the sample of 100 papers showed that 94% of the Japanese and English keyword pairs maintain good one-to-one correspondences. Based on the examination, we extracted a total of 112,364 Japanese and English keyword pairs (60,186 different ones) mechanically and use them as a basic bilingual keyword pairs. Table 1 shows some examples of the extracted keyword pairs.

Table 1. Japanese & English keywords

Japanese keywords	English keywords	frequency
キーワード	information retrieval	1
キーワード	keyword	39*
テキスト検索	information retrieval	1
テキスト検索	text retrieval	6*
テキスト検索	text search	3
検索指示語	keyword	1
広域情報検索	information retrieval	1
情報検索	information gathering	4
情報検索	information retreival	1
情報検索	information retrieval	320*
情報検索	information search	5
情報収集	information gathering	6*
情報収集	information retrieval	1
文献検索	bibliographic search	1
文献検索	document retrieval	11*
文書検索	document retrieval	9*
文書検索	text retrieval	1

**: The most general English translation is marked with '*' for each Japanese keyword.*

A comparison of dictionaries and bilingual keyword data shows that many of the keyword pairs are not listed in the dictionaries. The average number of translations is smaller in the standard dictionaries than that in the keyword data. This is partly because the dictionaries are prescriptive to some extent, while the keyword pairs reflect the authors' views. At the same time, it is also caused by the noises in the keyword data, *i.e.* errors of correspondence, related but improperly linked words, simple input errors, etc.

Table 2. Comparison of the technical dictionaries and the bilingual keyword data:

	dictionary terms	keywords pairs	common for both
Japanese words	20,636	37,170	3,966
English words	19,562	4,991	2,814
translation pairs	22,690	60,186	2,066
average number of translations per word (Jpn)	1.10	1.63	-
average number of translations per word (Eng)	1.16	1.21	-
maximum number of translations per word (Jpn)	7	86	-
maximum number of translations per word (Jpn)	6	29	-
number of English acronyms (Jpn)	844	1,007	212
number of English acronyms (Eng)	451	1,233	114
identical Japanese and English pairs	57	1,336	18

Dictionary terms are extracted from 4 dictionaries in the field of computer science (Aiso, 1993; Japan Society for Artificial Intelligence, 1990; Ralston, 1983; Shapiro, 1987).

Upon extraction from the basic keyword corpus, words from machine readable dictionaries can be integrated (with frequencies set to the infinite) to introduce more generality in the original corpus data.

2.2 Generation of Initial Keyword Graph

The initial graph expression is derived by representing Japanese and English keywords as nodes and their translation pairs as links. The frequency of the keyword pair appearing in the corpus is expressed as the capacity of the corresponding links. Figure 2 shows an example.

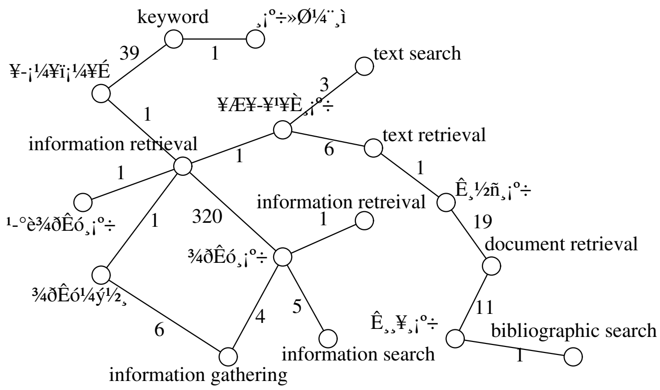


Figure 2. Initial keyword graph generated from keyword pairs in Table 1

The global keyword graph is composed of a number of disjoint sub-graphs, which we define as *bilingual keyword clusters*. Thus, by our definition, every node connected to each other is regarded as synonyms regardless of the capacity of the connecting links. In case of the example shown in Figure 2, all of the nodes belong to the same keyword cluster at the initial stage.

2.3 Partitioning Keyword Clusters

The initial keyword graph is partitioned according to the following procedures;

(1)Screening of Obvious Non-Errors

To reduce the computation cost, we discriminated obvious non-errors (unremovable links) as a simple pre-processing. Currently, this includes the recognition of identical keyword pairs, a frequency check, and the detection of minor examples. A given pair of terms are recognized as being identical when they have the same spellings in their notations, mostly English acronyms (see Table 2). 'Minor examples' refers to translation pairs which appear not so many times in the target corpus but not necessarily being incorrect, *i.e.* worth maintaining. Examples are: spelling errors (*e.g.* information retreival), variations, similar expression. Topologically, links with low frequency which connect non-significant keywords to significant clusters are considered to be minor examples. When a keyword has only one corresponding translation (*i.e.* the link itself), such link is maintained automatically as a minor example since it might be advantageous to maintain the expressional variations for IR applications .

(2) Detection of Possible Correspondence Errors : *edge cut*

The detection algorithm is based on a simple principle that *a set of links, which decompose a connected keyword cluster into disjoint sub-clusters when they are removed,* are the candidates of improper translations. In the conventional graph theory, such a link set is called an *edge cut* and the edge cut with the minimal total capacity is called a *minimum edge cut*. The minimum edge cut problem is one of the most principal problems in graph theory and a number of algorithms exist that guarantee sufficient performance for our purpose.

(3) Detection of Possible Homonyms : *node cut*

Though homonyms do not seem to occur frequently in a specific scientific domain, we have observed several such cases : < ATM (Asynchronous Transfer Mode)> and < ATM (Automatic Teller Machine)>. Possibly homonymous keywords can be detected utilizing the topological feature of the keyword cluster. It can be assumed that *homonymous nodes are the ones that decompose the cluster when the node and all the edges starting from the node are removed*. Thus, the problem is transformed again to the well-known *node cut problem* of the graph theory. Since most of the homonyms we have observed are abbreviations, we presently consider only acronyms as candidates for node cuts. However this may be insufficient in some cases.

Removing correct pairs inevitably causes oversplitting, i.e., generating more than one clusters with similar meanings. However, the appropriate size and level of semantic similarity in a cluster depends on the application. For example, the keyword pair <テキスト検索 (*text retrieval*), information retrieval> may be improper in view of strict terminological definition but not be incorrect for IR application.

2.4 Final Clustering Results

Figure 3 shows the example of the partitioning of the keyword cluster given in Figure 2. As a result of the detection of correspondence errors, three keyword pairs <キーワード (keyword), information retrieval>, <テキスト検索 (text retrieval), information retrieval>, <文書検索 (document retrieval), text retrieval>, are removed, and four clusters A, B, C, and D are newly created. The bold lines show that the links are marked as unremovable at the screening stage. It follows, that pairs such as <情報検索 (information retrieval), information retreival> (spelling error), <検索指示語 (search term), keyword> (rare case), and <広域情報検索 (wide-area information retrieval), information retrieval> (related but not equivalent pair) are retained even after the partitioning.

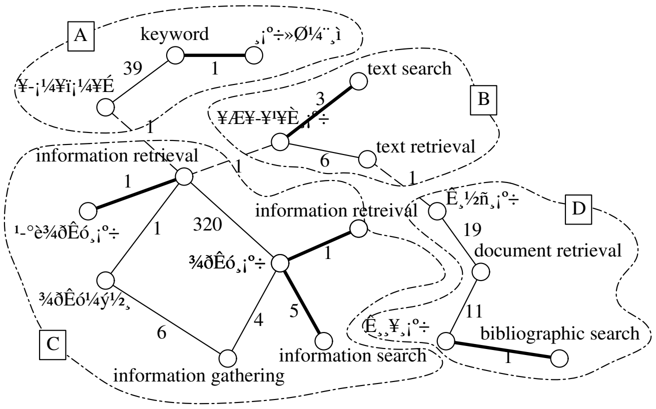


Figure 3: Example of partition of keyword cluster.

The next section reports the method and results of the search experiments using these bilingual keyword clusters.

3. Experiments

We examined the effectiveness of the clusters against two retrieval tasks: (1) CLIR: Japanese queries retrieving documents from an English collection (***J-E task***), and (2) Monolingual IR: Japanese queries retrieving documents from a Japanese collection (***J-JE task).*** In the following, we describe the collection used, indexing and query segmentation, term selection from the bilingual clusters, and retrieval results of both ***J-E*** and ***J-J*** tasks.

3.1 Collection

We used the test version of *NACSIS Test Collection 1, ver.0.3* (Kando, et al., 1998). The version of the collection contains more than 330, 000 documents, 30 search topics, and relevance assessment for each topic. Documents are abstracts of conference papers of the various subject domains and selected from *NACSIS Academic Conference Database* (NACSIS, 1997). More than half of the documents contains English-Japanese paired abstracts. For the ***J-J*** task, the Japanese fields in the document which contains Japanese abstracts (***J collection***, 338,668 documents), and the English fields in the document which contains English abstracts, were used for the ***J-E*** task (***E collection***, 186,809 documents) (See Table 3).

Table 3: Number of documents in the *NACSIS Test Collection 1 (test version 0.3)*

Sub collection	number of documents
JE collection (whole)	338,668
J collection	332,930
E collection	184,995

A topic consists of a title of the topic (1 - 3 words), a description (1 sentence) , a detailed narrative, and a list of concepts. Each narrative may contain detailed explanation of the topic, term definitions, background knowledge, purpose of the search, expected number of relevant documents, preference in text types, criteria of relevance judgement, and so on. We used only *description* fields in the experiments reported here. The topics consists of twenty-one Cross-lingual topics. Relevance assessments have three grades; relevant, partially relevant, and non-relevant.

3.2 Indexing, Query Segmentation, and Phrase Identification

Indexing and query segmentation are language dependent procedures. Therefore language-specific strategy is needed. Regarding document indexing, we indexed Japanese texts by character (uni-gram), and English texts by word. English terms appeared in Japanese texts were also indexed by word. Regardless of index types, each index entry maintains positional information (offset from the beginning of the database). It enable us to match word- or phrase-based query against strings longer than a character in the documents (Kando et al, 1998a).

Regarding query processing, phrase-based representations are used coupled with word-based representations (Kando et al., 1998a). Queries are Japanese natural language sentences. They were initially segmented into words using a Japanese morphological analyzer, *Chasen v1.5* (Matsumoto, et al., 1997). After discarding the stop phrases (*e.g.,* 文献 (document), 論文 (articles), について論じている (discussing on), が欲しい (I'd like to have, I'm interested in, etc.)), words and phrases are selected based on the patterns defined over part-of-speech tags. One pattern for phrase identification, for example, uses the maximal sequence of one or more adjectives followed by one or more nouns. All the constituent nouns in a phrase are also used as query terms.

3.3 Query Term Translation and Expansion using the Clusters

Each query term, both word and phrase, was translated or expanded using the bilingual keyword clusters described in Section 2. Each cluster may consist of more than one English and Japanese term pair and, in this paper, we treated all the terms in a cluster as synonyms. Therefore the clusters can be used to expand the query terms as well as to translate source-language query terms into target-language terms. We used the clusters to translate Japanese query terms into English in the J-E task. A cluster may contain more than one English term, thus, query terms were inherently expanded as well.

In order to examine the optimal partitioning of the keyword clusters and term selection criteria from the clusters for IR application, we tested these strategies listed below;

K3: keyword clusters partitioned by the minimum edge cut with capacity = 3, using all terms in the cluster,

KD3: keyword and dictionary term clusters partitioned by the minimum edge cut with capacity = 3, using all terms in the cluster,

KD10: keyword and dictionary term clusters partitioned by the minimum edge cut with capacity = 10, using all terms in the cluster

K3-2, K3-3, K3-5: From K3 clusters, select terms whose minimum edge cut value from the original query terms are no less than 2, 3, and 5, respectively, and

D3, D10: select terms originally listed in the dictionaries from KD3 and KD10 clusters, respectively.

We tested each of them in both the ***J-E*** and ***J-J*** tasks.

3.4 Search Engine and Search Formula Construction

The search engine is OpenText 6 SDK beta (OpenText Corp., Canada), which can handle both English and Japanese characters. The documents in the returned set are ranked using OpenText's "RankMode Relevance1". "Relevance1" ranks the members of the returned set based on the frequency of the terms in the document, the total number of index tokens in the document (document length), the number of the documents which contain the term (OpenText, 1997). This is a variant of *tf-idf,* which is used by most IR systems.

Weights were firstly calculated for each query unit, *i.e.* word or phrase, in the query, and the total weight was sum of weights of every constituent query units. Examples of weighting, phrase-construction, and synonyms are shown below;

BASE: $w(t_1)+w(t_2)+w(t_3)$

Phrase: $w(t_1)+w(t_2)+w(t_3)+w(t_2\ (3)\ t_3)$

Synonyms: $w(t_1)+w(t_2)+w(t_3 + t_{31} + t_{32} + t_{33} + t_{34} + t_{35})$

For example, a query sentence consists of 3 words of t_1, t_2, t_3, and weighting scores are calculated for each query unit as representing using ***w( )*** like **BASE**. When the sequence of "t_2. t_3" in a query is identified as a phrase, as shown as **Phrase,** the strings in the documents such as words t_2 and t_3 appear in this order but the maximum distance between the words is 3 characters in Japanese or 3 words in English match the query phrase.

When a word t_3 has synonyms of t_{31} . t_{32} . t_{33}. t_{34}, t_{35}, all the synonyms are wrapped by a pair of parentless as shown as **Synonyms**. all words within is as occurrences of a single pseudo-term whose document frequency (*df*) is the sum of *df*'s for each word in the parentheses. This synonyms operation de-emphasizes infrequent words and has a effect of disambiguation. When a query is expanded using the keyword clusters, some of the expanded terms can be very infrequent ones since the our approach maintain less frequent but important terms in the clusters. A document with an infrequent term receives an extraordinarily high weight and it may lead the retrieval into wrong direction. Therefore de-emphasize the effect of such infrequent or rare terms are necessary.

TREC evaluation program is used to calculate the recall and precision. Non-interpolated average precision is used as the basis of evaluation. The paired sign test is used as a significant test.

4 Retrieval Results

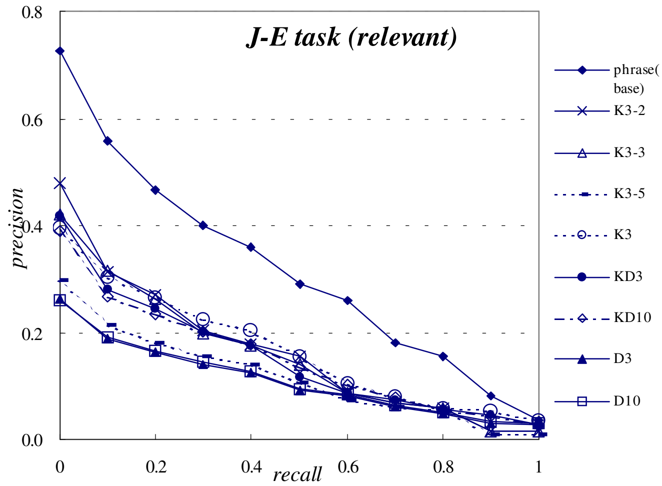
Table 4 shows the average number of the translated and expanded terms using the keyword clusters for a query term. In the J-J task, expanded terms include both Japanese and English since English terms can be matched to English terms in English texts and the ones appearing in Japanese texts.

	K3	KD3	KD10	K3-2	K3-3	K3-5	D3	D10
J-E	6.81	4.48	3.66	1.34	0.87	0.62	0.84	0.77
J-J	10.7	7.01	5.72	1.8	1.12	0.72	1.4	1.3

Table 4: Average number of the expanded terms

4.1 CLIR Results (*J-E task)*

The results are shown in Figures 4 and 5. Terms in the twenty-one queries were translated into English terms using the keyword clusters, and were tested against the *E-collection.* The baseline was monolingual retrieval using



	K3-2	K3-3	K3-5	K3	KD3	KD10	D3	D10
at 0.00	66.0%	57.9%	41.0%	54.4%	57.5%	53.8%	36.0%	35.9%
at 0.10	56.4%	56.8%	38.3%	54.6%	50.1%	48.2%	34.0%	34.5%
at 0.20	57.9%	56.3%	39.1%	56.8%	52.6%	50.1%	35.0%	35.4%
at 0.30	51.2%	49.9%	38.7%	56.1%	50.6%	50.4%	35.1%	36.5%
at 0.40	50.1%	48.9%	39.1%	56.5%	48.8%	49.5%	34.7%	35.9%
at 0.50	53.9%	48.3%	36.9%	53.6%	40.1%	46.4%	31.2%	32.1%
at 0.60	33.1%	33.2%	28.4%	40.7%	33.0%	39.5%	31.3%	32.3%
at 0.70	37.7%	37.7%	34.2%	43.4%	41.2%	43.7%	34.0%	35.4%
at 0.80	37.5%	37.5%	34.3%	37.0%	36.6%	36.0%	31.1%	32.7%
at 0.90	17.8%	17.8%	14.0%	65.3%	55.7%	54.4%	38.5%	41.6%
at 1.00	39.7%	39.7%	31.4%	101.6%	79.9%	77.4%	74.9%	81.9%
Average*1	49.3%	48.4%	35.8%	52.4%	47.0%	46.4%	33.5%	34.8%

phrase identification and the same as the one of the *J-J task*.

Figure 4: Retrieval results with relevance assessment with the level of "Relevant" in the J-E task : the table show the ***% of the baseline*** (monolingual retrieval using phrase). *1: Average precision (non-linterpolated) for all relevant documents.

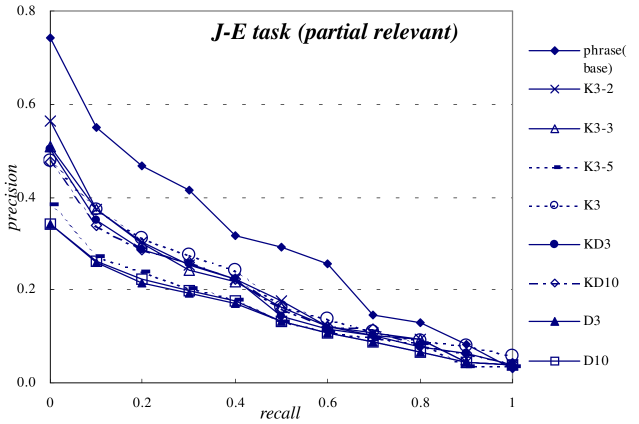


Figure 5: Retrieval results with relevance assessment

	K3-2	K3-3	K3-5	K3	KD3	KD10	D3	D10
at 0.00	76.1%	68.4%	51.9%	64.5%	67.9%	64.2%	46.0%	46.0%
at 0.10	67.8%	67.7%	49.3%	67.7%	63.5%	61.3%	46.8%	47.6%
at 0.20	64.8%	64.3%	50.9%	66.3%	61.0%	61.4%	45.9%	47.3%
at 0.30	61.3%	58.3%	49.4%	66.4%	62.0%	62.4%	46.7%	48.1%
at 0.40	70.2%	68.5%	56.8%	76.9%	70.2%	70.7%	54.4%	55.8%
at 0.50	60.3%	55.4%	44.9%	55.6%	48.7%	54.9%	45.3%	45.3%
at 0.60	47.0%	47.1%	43.1%	53.4%	45.1%	47.1%	41.9%	41.9%
at 0.70	72.7%	70.0%	65.5%	76.1%	70.1%	78.1%	59.5%	59.5%
at 0.80	71.4%	71.4%	66.5%	69.6%	59.3%	61.4%	51.2%	51.2%
at 0.90	53.4%	53.4%	43.5%	99.0%	76.3%	78.8%	52.4%	52.4%
at 1.00	126.8%	128.2%	116.8%	193.3%	131.2%	135.9%	124.5%	124.5%
Average*1	62.0%	61.3%	49.8%	65.7%	59.1%	60.0%	46.1%	47.0%

with the level of "Relevant + Partial Relevant" in the J-E task: the table show the ***% of the baseline*** (monolingual retrieval using phrase). *1: Average precision (non-linterpolated) for all relevant documents.

The tables show the relative performance of each strategy compared to the baseline of the monolingual retrieval. The results showed that cross-lingual retrieval effectiveness of J-E task achieved 52.4% below that of monolingual retrieval at the average precision for relevant level, and 65.7% for partial relevant level. It has been reported that the CLIR effectiveness using machine readable dictionaries can be 40-60% below that of monolingual retrieval (Bellasterors & Croft, 1997) and the results of the latest TREC's Cross-language track are 50-75% (Schauel & Sheridan, 1998). Compared to those standards, our results using automatically constructed keyword clusters are fairly good as a start point.

The number of translated terms per a original query term seems an important parameter to estimate the contribution of the clusters to the search effectiveness. The drop of the effectiveness relative to the monolingual tended to large for the strategies with fewer expanded terms. For D3 and D10, the dictionary terms in the clusters were the poorest in this J-E task, whereas these achieved fairly well in the J-J task reported later. On the other hand, K3-2 and K3-5 were rather well achieved with small number of translated terms.

When all terms in the clusters were used like K3,KD3, KD10, the effectiveness improved. Proper term selection from the cluster based on the topological feature of the clusters like K3-2, K3-3 achieved also well, especially at low recall (high precision). However, too many reduction from the cluster both based on topology like K3-5 and based on term-sources (e.g. dictionary terms) like D3,D10 achieved rather poor in CLIR.

4.1.1 Failure analysis:

Decline of the effectiveness was mainly caused by translation errors of query terms. There are two types of errors:

- **Lack of translation**: a query term may not be found in any keyword clusters, then the term can not be translated.
- **Extraneous terms**: the clusters may contain several terms, and some of them may be not suitable for the context of the query.

The first type error: lack of translation:

The first type of error depends on the available resources for translation (i.e. bilingual dictionaries, parallel corpora, etc.) and the scope of them. In our experiments, two topics among total twenty-one have no or only one query term can be translated. The effectiveness for these two topics were extremely bad, retrieved no or only few target language relevant documents.

Table 5: Average number of query terms and lack of translations per a query

	K3-2	K3-3	K3-5	K3	KD3	KD10	D3	D10
Qwords with translation	54	44	41	66	67	67	53	51
Qphrases with translation	19	17	15	22	22	22	10	11
total Q terms	73	61	56	88	89	89	63	62
% of Qterms with translation	62.9%	52.6%	48.3%	75.9%	76.7%	76.7%	54.3%	54.3%
maximum % per a query	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
minimum % per a query	16.7%	0.0%	0.0%	16.7%	16.7%	16.7%	0.0%	0.0%

It is also effected by the strategy of cluster. Table 5 shows that when all the terms in the clusters were used like K3, KD3, KD10, more than 75% of the total query terms can be translated. Others can translate about half of the query terms. However, when selecting terms from clusters based on the topological feature like K3-2 and K3-3, more phrases can be translated than the dictionary terms clusters of D3 and D10. As a unit of information, a phrase is more content-bearing and precise than the sum of its constituent words, thus phrase-based representation can lead to improvements in retrieval effectiveness. The retrieval results also suggested the effects of phrases in CLIR: the strategy which can translate more phrases in the source language queries into target language terms showed higher effectiveness.

This result indicates that using all terms in a cluster (*i.e.*, K3, KD3, KD10) or proper term selection based on the topological feature of the clusters (*i.e.,* minimal edge cut is 2 or 3 (K3-2 or

K3-3)) leads to diminish the first type errors, i.e., lack of translation, and especially have an advantage of phrase translation.

Second type error: extraneous terms:

Detailed examination of retrieval results of each query revealed that some of the clusters are extraordinarily big and translation via such clusters may introduce terms which are inappropriate to the context of the query. The examples are shown below:

Examples of English terms expanded via K3 clusters: *Only underlined terms are included in K3-2 clusters:*

解析 (analysis) --> analysis, traffic analysis, analyze, theoretical performance analysis

[Query: Are there any research concerning the automatic analysis of complex nouns using both statistical and symbolic method together**?]**

排他制御 (exclusive control) --> parallel processing, parallel computer, multiprocessor, multi processor, parallel, concurrency control, massively parallel, parallelization, massively parallel computer, parallel machine, parallel execution, concurrent processing, parallel computation, parallel computing, multi processor system, concurrency, massively parallel processing, parallelism, mutual exclusion, concurrent process, massively parallel processor, highly parallel, parallel process, parallel processing system, multi process, multi processors, selection, exclusive control, future, ...

[Query: exclusive control for simultaneous sending to buses**]**

In the first example, a word "**解析**(analysis)" can be used in the context of network traffic analysis, but the context of the query is "complex noun analysis" therefore "traffic analysis" is extraneous term for the query. In the second example, a word "**排他制御** (exclusive control)" is somehow related to the concept of "parallel computing" and both concepts can be appeared in many documents however the context of the query is "bus control". "Parallel computing" is less related to it and rather extraneous to the context of the query. Extraneous terms like them can be seen in K3 and KD3 clusters, but hardly found in K3-2, K3-3, K3-5, D3 and D10 clusters. In these cases, term selection from clusters using topological feature like K3-2, K3-3, K3-5 seems effective to discard such extraneous terms and lead to improvements of CLIR effectiveness.

Based on the discussion of these two types of translation errors, K3-2 and K3-3, selecting terms from keywords clusters based on topological features, are promising to de-emphasize both types of errors.

4.2 Monolingual Retrieval (J-J task)

The results of monolingual retrieval of J-J task using keyword clusters to expand queries are shown Tables 6 and 7. Query expansion using keyword clusters improved the retrieval effectiveness of J-J task 13.4% over the baseline at the average precision for relevant level, and 14.2% for partial relevant level. The baseline is the monolingual retrieval using automatic phrase identification and phrase weighting. It is better than word-based or n-gram phrase processing in our previous reports (Kando et al, 1998a)

Again the number of expanded terms seems an important parameter to estimate the contribution to improve the search effectiveness. However, dictionary term clusters of D3 and D10 also improved effectiveness fairy well in low recall (high precision) as shown in Table 6.

The extraneous terms is also problematic for bigger clusters like K3 and KD3. Futher study is needed ot optimize the term selection from the clusters.

Figure 6. Average precision and precision at low recall with relevance assessment with the level of "relevant" in J-J task

	phrase(base)	K3-2	K3-3	K3-5	K3	KD3	KD10	D3	D10
Average*1	0.299	0.334	0.336	0.320	0.339	0.327	0.329	0.321	0.335
%improved over the base		11.4%	12.0%	7.0%	13.2%	9.3%	9.8%	7.3%	11.8%
At 5 docs	0.4328	0.5524	0.5524	0.5429	0.5619	0.5333	0.5333	0.5714	0.5714
At 10 docs	0.4370	0.4857	0.4810	0.4762	0.4857	0.4619	0.4810	0.4667	0.4762
At 15 docs	0.3970	0.4444	0.4508	0.4413	0.4413	0.4222	0.4349	0.4317	0.4349
At 20 docs	0.3671	0.4095	0.4095	0.4024	0.4214	0.4048	0.4143	0.3976	0.4071
At 30 docs	0.3314	0.3540	0.3603	0.3571	0.3524	0.3524	0.3635	0.3413	0.3556

Table 7. Average precision and precision at low recall with relevance assessment with the level of "relevant+partial relevant" in J-J task

	phrase(base)	K3-2	K3-3	K3-5	K3	KD3	KD10	D3	D10
Average*1	0.297	0.334	0.335	0.321	0.339	0.327	0.340	0.311	0.325
%improved over the base		12.3%	12.6%	8.1%	14.0%	10.1%	14.2%	4.8%	9.2%

5. Discussion

We have proposed an approach to CLIR using automatically generated keyword clusters based on graph- theoretic methods. There have been many studies on CLIR in recent years. Methodologically, these can be categorized into three: dictionary-based methods (Ballesteros & Croft, 1997; Ballesteros & Croft, 1996); corpus-based methods (Davis & Dunning, 1993; Landauer & Littman, 1990; Carbonell, 1997); and machine translation techniques (Collier et al, 1998).

Statistical corpus-based methods are also predominant in the automatic thesaurus construction and automatic identification of bilingual lexical pairs. On the other hand, in statistical approach, the method totally depends on the frequency of the co-occurrence of the terms, and, thus, "***less frequent but important***" examples tend to be discarded.

One of the problems of the CLIR is, despite promising experimental results with each of these approaches, each approaches has drawbacks associated with the availability of resources (Ballesteros & Croft, 1998; Schaubel & Sheridan, 1998). For dictionary-based methods, the coverage of dictionaries or thesauri are often not sufficiently broad and deep, thus domain specific terms or new concepts, which are critical for retrieval of technical and scientific documents used here, tend not to be listed. Lack of resources is also problematic in corpus-based methods; parallel corpora are not always readily available. Regarding machine translation techniques, we can not ignore the cost of linguistic analysis.

In the contrast to these, the approach reported here has several advantages: Regarding to the resources, the keyword data we used here has also advantages, *i.e.*, *subject-spe*cific bilingual keyword corpora are *readily available* in machine-readable form for a great many subject domains, readily segmented into terms, and well aliened (albeit with some noise). Therefore, they are rather easy to handle for our purpose. In addition to this, the graph-theoretic approach as advantages: (1) by utilizing topological features of the graph, low-frequency keywords can be treated properly and are usable in IR; (2) the clusters contain not only J-E pairs, but also J-J and E-E pairs, and (3) this was achieved with reasonable computational cost.

Our approach for the CLIR using keyword clusters based on only less than 10% of database as a corpus achieved the effectiveness of 52.4% of monolingual at average precision for all relevant documents at "relevant" level and 65.7% for :relevnat+partial relevant" level with automatic phrase identification and synonym processing, without any manual interaction. This is comparable to existing researches and fairy well as a start point. However this still leaves a lot of room for improvement both in the aspect of IR and clustering.

For the first type of translation error, i.e., lack of translation, these directions are possible: adding keywords corpus or other lexical resources as a corpus of clustering, introducing query expansion both before and after query expansion (Bellasteros & Croft, 1997). Pre-translation query expansion of Local Context Analysis is reported to provide more solid basis of the translation and improves recall, and post-translation expansion lead to improvement of precision. Our approach included a kind of query expansion using keyword cluster, but do not use context of document nor query. For the relevancy of the retrieved documents, context of both document and query are critical. Therefore, introducing automatic query expansion using context of documents is expected to reduce the both types of translation errors.

The problems caused by the second type of error can be thought as the problems of disambiguation in IR: phrase searching, synonym operators, query expansion both pre- and post-translation, co-occurrence check using the context of the documents are reported to effective for the disambiguation in IR (Bellasteros & Croft 1998). We already incorporated phrase searching and synonyms, enhancement of phrase searching (Kando et al, 1998a), query expansion, and disambiguation using the context will be the direction for the further studies.

In IR with Japanese language, local feedback, i.e. extracting high-frequency terms from relevant texts, or text that are supposed to be relevant, can be expensive. The reason for this is that there is no explicit word separators in Japanese text, and simple character-based query segmentation is not always as effective as word-based query segmentation in our retrieval experiments. The author-assigned keywords that we used in this study are already segmented into terms, thus, local feedback utilizng these keyword fields will be one of the practical approaches. It is difficult to harmonize the evidences from different fields, such as text and manually assigned keywords, because they are quite different; few studies have tried the retrieval based on the combination of content and manually assigned keywords.

Regarding the future work in clustering, incorporating NLP technique to refine the procedure is an interesting extension. One possible application of NLP technique is the detection of unremovable links and homonyms. The correspondence between components of compound terms is also important aspect in dealing with different languages.

ACKNOWLEDGMENTS

This research is supported by "Research for the Future" Program JSPS-RFTF96P00602 of the Japan Society for the Promotion of Science.

REFERENCES

Aizawa, A. ; Kageura, K. 1998a, "An approach to the automatic generation of multilingual keyword clusters". COMPUTERM '98, Aug. 1988. Montreal, Canada

Ballesteros, L.; Croft W. B. 1996a. Dictionary-based methods for cross-lingual information retrieval. In Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, p.791-801.

Ballesteros, L.; Croft.W.B. 1996b. Statistical methods for cross-lingual information. a paper presented in the Workshop on Cross-Linguistic Information Retrieval at the 19th Annual International ACM SIGIR Conference of Research and Development in Information Retrieval, Zurich, Switzerland.

Ballesteros, L.; Croft.W.B. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In Proceedings of the 20th Annual International ACM SIGIR Conference of Research and Development in Information Retrieval, Philadelphia, Pennsylvania, USA, p.84-91.

Ballessteros, L.; Croft, W.B. 1998. Resolving ambiguity for cross-language retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference of Research and Development in Information Retrieval, Melbourne, Australia, p.64-71

Carbonell, J.G. ; Yang, Y.' Frederking, R.E.; Brown, R.D.; Geng, Y.; Lee, D. 1997, Translingual information retrieval: a comparative evaluation. In 15th International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan, August 23-29, 1997, p.708-714

Collier, N.; Hirakawa, H.; Kumano, A. 1998. Cross language information retrieval' an experiment in bilingual news article alignment from the internet using MT. JSPS-HITACH Workshop on New Challenges in Natural Language Processing and its Application ' Integration of linguistics-based and corpus-based approaches, May 26-28, 1998, Tokyo, p.133-137

Dunning, T.; Davis, M. 1993, Multi-lingual information retrieval. Technical report MCCS-93-252, Computer Research Laboratory, New Mexico State University, 1993

Kando, N. 1997, Cross-linguistic scholarly information transfer and database services in Japan. a paper presented at the Annual Meeting of the American Society for Information Science, Nov. 2-7, 1997, Washington, D.C., U.S.A.,

Knado, N.; Kageura, K.; Yoshioka, M.; Oyama, K. 1998a, Phrase processing methods for Japanese text retrieval. ACM-SIGIR'98 Workshop of Information Retireval: Theory into Practice, August 28, 1998, Melbourne, Australia, p.13-19.

Kando, N. ; Koyama, T.; Oyama, K.; Kageura, K.; Yoshioka, M.; Nozue, T.; Matsumura, A.; Kuriyama, K., 1998b, NTCIR: NACSIS Test Collection Project" [poster]. the 20th Annual BCS-IRSG Colloquium on Information Retrieval Research, March 25-27, 1998, Autrans, France.

Landauer, T.K.; Littman, M.L. Fully automatic cross-language document retrieval. In Proceedings of the Sixth Conference on Electronic Text Research, p.31-38, 1990.

Matsumoto, Y. et al. 1997, Japanese Morphological Analyzer Chasen 1.5, NAIST, 1997.

NACSIS, 1997, Introduction to the National Center for Science Information Systems, NACSIS, (http://www.nacsis.ac.jp/)

Myaeng, S.H.; Jeong, K.S.; Kwon, Y.H., 1997, The effect of a proper handling of foreign and English words in retrieving Korean text. In proceedings of the 2nd International Workshop on Information Retrieval with Asian Languages, Oct., 8-9, Tsukuba, Japan

Shauble, P.; Sheridan, P. (1998) "Cross-language Information Retrieval (CLIR) Track overview" In Proceedings of TREC6, p.25-30.

Xu,Jinxi. Croft,B., 1996, Query expansion using local and global document analysis. In the proceedings of the 19th Annual International ACM SIGIR Conference of Research and Development in Information Retrieval, Zurich, Switzerland, p.4-11