Segmented LSI for Fully Automated Large-scale Cross-Language Information Retrieval

Tatsunori Mori
Graduate School of Environment and Information Sciences
Yokohama National University
79-5 Tokiwadai, Hodogaya, Yokohama 240-8501, JAPAN
mori@forest.eis.ynu.ac.jp

Abstract

In this paper, we propose a method to apply Cross-Language Latent Semantic Indexing(CL-LSI), which is a fully automated scheme for Cross-Language Information Retrieval, to a large-scale bilingual corpus. When we construct one monolithic word space of LSI with a large-scale corpus, we encounter problems not only the increase in ambiguity of word translation, but also the difficulty in singular value decomposition(SVD), which is the essential process of LSI. To cope with problems, we introduce a new LSI method in which a large bilingual corpus is divided into smaller sub-corpora enough for system to apply SVD to them. Each document is placed into one of sub-spaces which is made from the sub-corpus most similar to the document. In the search, queries are placed into every sub-space, and similarity between them are calculated. We show that it is important to adjust similarity calculation according to unknown words in each word sub-space.

1 Introduction

Drastic growth of the Internet increases the need of seamless access to multi-language documents. Cross-Language Information Retrieval (CLIR) is one of the most important technologies to satisfy the need.

As described in literatures(Oard and Dorr, 1996; Hull and Grefenstette, 1996; Kikui, 2000), many kinds of methods of CLIR make use of trans-lingual dictionaries. Those dictionaries are usually compiled and examined by hand. The accuracy of translation of those systems should be high, although the accuracy depends on not only the scale and quality of the dictionaries but also the way to use them.

On the other hand, there are several methods which extract trans-lingual information from multi-lingual resources such as bilingual corpora, and use the resouces for CLIR. Those methods are very attractive because of their automated scheme. Some of them are based on automatic construction of trans-lingual dictionaries. Other schemes utilize such resources to map documents into their surrogates such as document vectors. Cross-Language Latent Semantic Indexing(CL-LSI) is one of the representative schemes of the latter.

Although we should also take account of the cost to compile and maintain such bilingual corpora, some kinds of multi-lingual corpora are growing larger year by year without considerable labor, like summaries of technical papers written in more than one language.

Of course, the corpus-based systems is generally supposed to be less effective than systems with dictionaries compiled by hand. Against the expectation, Carbonell et al. (Carbonell et al., 1997) shows, by middle-scale experiments with 1134 dual-language documents, that an example-based Machine Translation establishing corpus-based term equivalences is the most effective, and both CL-LSI and the generalized vector space model follow it, but the Machine-Readable-Dictionary-based query translation is less effective than the others.

However, it still is not clear how precisely (or inaccurately) the systems such as CL-LSI can perform CLIR for a large-scale document database without dictionaries. Moreover, CL-LSI has the fatal problem that it could not treat a large-scale document database. When the target of information retrieval is a large-scale document set, we have to extract trans-lingual information from the a large-scale bilingual corpus, which is comparable to the target document set in terms of scale and coverage of domains. The trans-lingual information of CL-LSI is obtained by the singular value decomposition(SVD) of a term-document matrix. The SVD process for a large-scale term-document matrix will break down because of shortage of computers' main memory.

In this paper, we introduce a new LSI method, called *Segmented LSI*, in which a large bilingual corpus is divided into smaller sub-corpora enough for system to apply SVD to them. Each document is placed into one of sub-spaces which is made from the sub-corpus most similar to the document. In the search, queries are placed into every sub-space, and similarity between them are calculated.

Through the evaluation of NTCIR2, we demonstrate that our scheme of LSI can treat a large-scale corpus. We show that it is important for segemented LSI to adjust similarity calculation according to unknown words in each word subspace. In order to compare our segmented LSI with the original, monolithic LSI, we also conduct another experiment with a middle-scaled corpus, which is manageable with the original LSI. The result shows that our segmented LSI has almost same or a little bit higher effectiveness than the original LSI.

2 Cross Language Latent Semantic Indexing

Cross-language LSI (CL-LSI) is a fully automatic method for cross-language document retrieval in which no query translation is required (Dumais et al., 1996; Dumais et al., 1997). Queries in one language can retrieve documents in other languages as well as the original language.

2.1 Latent Semantic Indexing

A central theme of LSI is that term-term interrelationships can be automatically modeled and used to improve retrieval(Deerwester et al., 1990). This is critical in cross-language retrieval since direct term matching is of little use. LSI examines the similarity of the "contexts" in which words appear, and creates a reduced-dimension feature space in which words that occur in similar contexts are near each other. LSI uses a method from linear algebra, singular value decomposition (SVD)(Leach,), to discover the important associative relationships. The learned associations are specific to the domain of interest, and are derived completely automatically. For information retrieval we begin with a large term-document matrix. The (i, j) element of the matrix is the frequency of the term i in the document j. This term document matrix is decomposed into a set of k, typically 200-300, orthogonal factors from which the original matrix can be approximated by linear combination. This analysis reveals the "latent" structure in the matrix that is obscured by variability in word usage.

Traditional vector methods represent documents as linear combinations of orthogonal terms. In contrast, LSI represents terms as continuous values on each of the orthogonal indexing dimensions. Terms are not independent. When two terms are used in similar contexts (documents), they will have similar vectors in the reduced-dimension LSI representation. LSI partially overcomes some of the deficiencies of assuming independence of words, and provides away of dealing with synonymy automatically without the need for a manually constructed thesaurus. The result of the SVD is a set of vectors representing the location of each term and document in the reduced k-dimension LSI representation. Retrieval proceeds by using the terms in a query to identify a point in the space. Technically, the query is located at the weighted vector sum of its constituent terms. Documents are then ranked by their similarity to the query, typically using a cosine measure of similarity.

New documents (or terms) can be added to the LSI representation using a procedure we call "folding in". This method assumes that the LSI space is a reasonable characterization of the important underlying dimensions of similarity, and that new items can be described in terms of the existing dimensions. A document is located at the weighted vector sum of its constituent terms.

2.2 Cross-Language Retrieval Using LSI

LSI could easily be adapted to cross-language retrieval as shown in Figure 1. An initial sample of documents is translated by human or, perhaps, by machine, to create a set of dual-language training documents.

A set of training documents is analyzed using LSI, and the result is a reduced dimension semantic space in which related terms are near each other. Because the training documents contain both terms of two languages, the LSI space will contain terms from both languages, and the training documents. This is what makes it possible for the CL-LSI method to avoid query or document translation. Words that are consistently paired will be given identical representations in the LSI space, whereas words that are frequently associated with one another will be given similar representations.

The next step in the CL-LSI method is to add (or "fold in") documents in just one language. This is done by locating a new document at the weighted vector sum of its constituent terms. The result of this process is that each document in the database, whether it is in one of two language, has a language-independent representation in terms of numerical vectors. Users can now pose queries in one of those languages and get back the most similar documents regardless of language.

3 Issues in making LSI spaces from a huge set of dual-language documents

The CL-LSI can be considered as the method which is effective mainly for document database in a certain specific domain. If the database includes documents from diverse domains, we have to collect a large number of dual-language documents in order to make a huge LSI space which has enough vocabulary for the document database. When we would like to obtain such an LSI space, we face the problem in the process of SVD. Since SVD is a kind of operation for matrices, the time and space complexity of computation will increase for larger data. Thus, if we use a huge set of dual-language documents, the process will break down because of the shortage of computers' main memory.

For example, we can find about 180 thousand dual-language summaries in the NTCIR1 corpus and 370 thousand words in it(NTCIR Project, 2000). The document-word matrix for the summaries will have 67G elements. It can not be

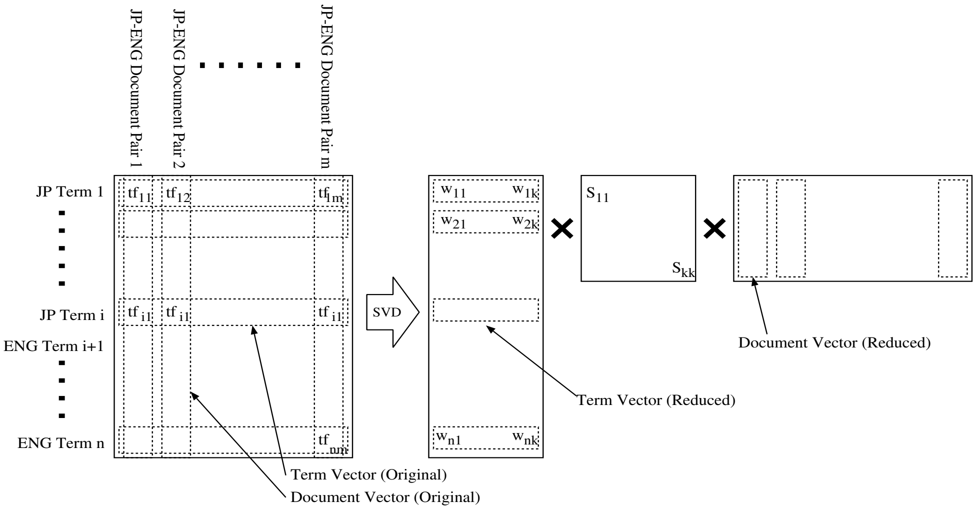

Figure 1: Cross-Language Latent Semantic Indexing

stored in the memory of computers except for super computers, even if the matrix is rather sparse.

Therefore, we introduce a method in which the large bilingual corpus for training is divided into smaller sub-corpora according to both the similarity among documents and computers' resource. Figure 2 shows the overview of our scheme. We expect the introduction of the multiple LSI subspaces to contribute toward the following objectives.

- SVD can be performed to make LSI spaces.
- The ambiguity in translation will be decreased if the area associated with each sub LSI space is appropriately restricted.

In order to adopt the method, we have to consider the following issues:

- How can we divide a corpus into sub corpora?
- How can we place (new and mono-lingual) documents in the set of LSI spaces.
- How can we retrieve the documents in the set of LSI spaces.

In the next section, we will describe our approach to these issues.

4 Segmented LSI for a huge set of dual-language documents

4.1 Dividing traing corpus into sub-corpora

In our approach, a huge set of training corpus is segmented into several sub-corpora in order for computer to perform SVD. If each sub-corpus is limited to a certain area, the context in the documents will be restricted and the variations of translation of words would be also decreased. Therefore, it would be effective in CLIR to divide a training corpus according to the similarity among documents.

Although clustering algorithms are usually used to do that, it costs a huge amount of computational resources to apply an ordinary clustering algorithm to a huge document set[1]. We also have to adjust the size of subsets of documents according to computational resources, even if a certain clustering algorithm can be used.

Thus, we utilize anothor method in our scheme. In real situations, documents are usually accompanied with some useful information to guess the areas of documents. For example, each technical paper includes some information about 'area name' like the name of society. Therefore, in this paper we adopt the following approximated way of clustering in which the dual-language documents of the same area name treated as one document group and some of document groups are merged or divided according to the limit size of document group.

1. Classify dual-language documents into area groups according to the area name of each document.
2. For each dual-language document, make a tfidf-based document vector $(tfidf_1, \ldots, tfidf_i, \ldots, tfidf_n)$, where the weight $tfidf_i$ is the TFIDF value of the term i in the document.
3. For each area group, calculate the 'area vector' by averaging all of the document vector in the group.

[1] Recently, new approximation algorithms like BIRCH have been proposed for clustering a huge data set.

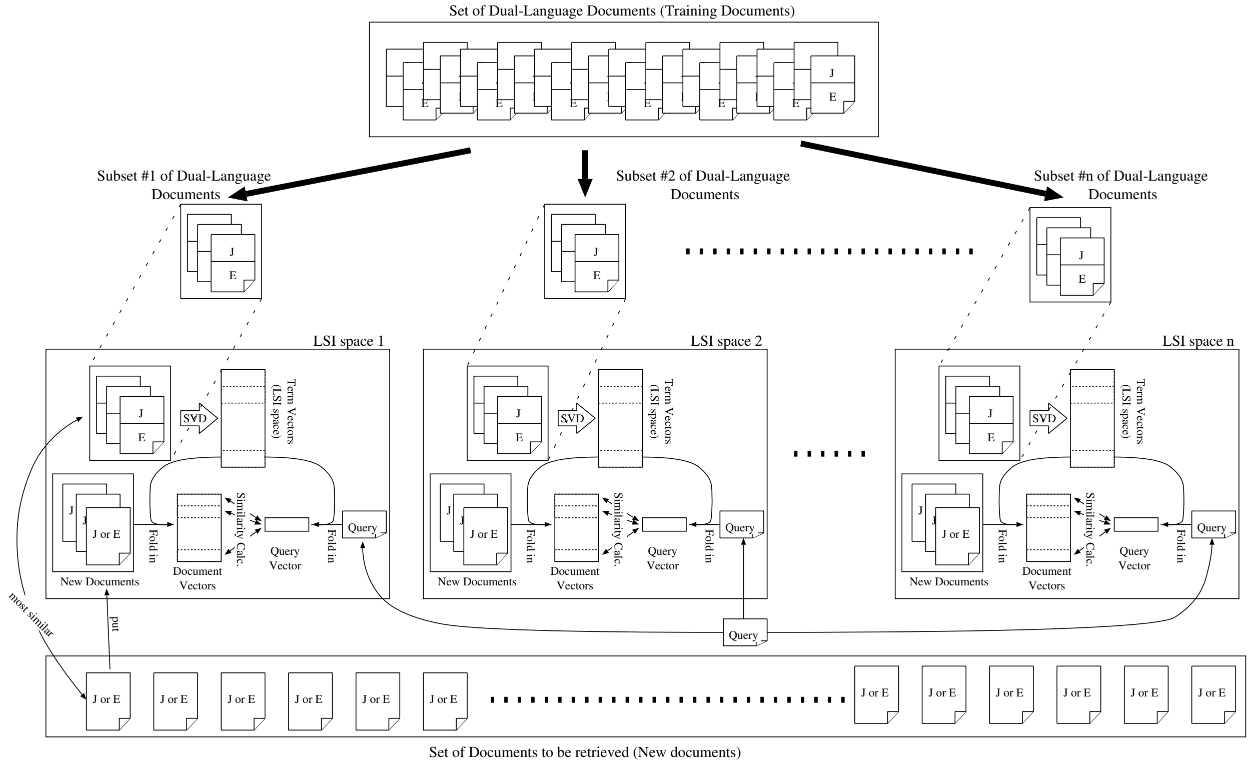

Figure 2: Proposed Scheme — Segmented Cross Language LSI

4. Select several large area groups manually. We call the groups 'major area groups'.

5. For each of other area groups, find the most similar major area group and marge it to the major area group according to similarity of vectors. The cosine correlation is adopted as the similarity of vectors.

6. For each area group the size of which exceeds a certain limit, divide it into sub groups of the required size. The limit is determined according to computational resources.

7. Update the area vectors of existing area groups.

4.2 Storing Documents

In the scheme of CL-LSI, the (mono-lingual) documents to be retrieved are different from the dual-lingual documents which is used to make the LSI space. Thus, we have to "fold in" all of mono-lingual documents to the LSI space by using term vectors. Since we have plural LSI spaces corresponding to area groups, the structure of document vectors depends on which of LSI spaces is selected.

When we have plural LSI spaces, several ways would be supposed to make document vectors as follows.

(a) Each document is placed in every LSI space as different vectors.

(b) Each document is placed in one selected LSI space.

Since all of translation information is took into account, the method (a) would be expected to be more effective than the method (b). The method (a), however, requires a huge storage system because each LSI space has the full set of document vectors. Accordingly, we adopt the realistic method (b).

In (b), we have to consider the way to select one 'suitable' LSI space for each document. Because of accuracy of translation, it is desirable that a document is put into the LSI space which made from (training) documents in the same area as the target. Therefore, each document is placed to the LSI space the area vector of which is most similar to the tfidf-based vector of the document.

In order to "fold in" the mono-lingual documents to the LSI spaces, we use the following formula[2].

$$\mathbf{D} = \sum_{T_i \in D} tf(T_i, D) idf(T_i) \mathbf{T}_i, \qquad (1)$$

where

[2] It is different from the original LSI because we use IDF values.

$\mathbf{D}$:	Document vector of document D in the LSI space.
$tf(T_i, D)$:	Frequency of term T_i in document D.
$idf(T_i)$:	IDF value of term T_i, $\log \frac{N}{df(T_i)} + 1$, where $df(T_i)$ is the document frequency of T_i in database.
$\mathbf{T}_i$:	Term vector of term T_i in LSI space.

4.3 Document Retrieval from Plural LSI Spaces

In the CL-LSI method, each query is also represented as a vector in the LSI space. According to the similarity between the query vector and each document vector, all documents are ranked in terms of query. The retrieval of documents is performed based on the ranking information.

Since we have plural LSI spaces in our scheme, we retrieve the documents by the following procedure:

1. Make one query vector for each LSI space by (1), in order to compare the query with all of documents.
2. In each LSI space, calculate the similarity between the query vector and each document vector.
3. Rank all documents in all LSI spaces according to their similarity.

5 Problem of Unknown Words arising from Dividing Bilingual Corpus

In the CL-LSI method, the unknown words, which do not appear in the set of dual-language documents, are totally ignored because we cannot obtain the translation information of them. Thus, the accuracy of retrieval will be degraded when there are a number of words which appear not in dual-language documents but in documents to be retrieved. It is an inevitable problem because of methodology of LSI.

Unfortunately, we have another unknown-words problem in our scheme. It is caused by the division of corpus. When we divide a training corpus, there may be words which do appear not in some sub-corpora but in the other sub-corpora. Since each LSI space has a different set of unknown words from others, in some cases we can not obtain desired results in document retrieval.

For instance, let us consider the case where with the query $Q(T_a, T_b, T_c)$ of three terms T_a, T_b and T_c the system retrieves documents in the LSI spaces TS_1 and TS_2. We suppose that the space TS_1 has T_a but does not have T_b and T_c, and the document $D_1(T_a)$ with T_a is placed into the space. On the other hand, we also suppose that the space TS_2 has T_a, T_b and T_c, and the document $D_2(T_a, T_b)$ with T_a and T_b is placed into the space.

In this situation, the document $D_2(T_a, T_b)$ is more preferable to $D_1(T_a)$ as a retrieved document for the query, and we expect that the similarity between $D_2(T_a, T_b)$ and the query is larger than the similarity between $D_1(T_a)$ and the query. However, in reality the similarity about $D_1(T_a)$ is larger than that about $D_2(T_a, T_b)$. The reason is as follows. Since only T_a is considered in the process of the similarity calculation in TS_1, the query is substantially regarded as T_a and consequently $D_1(T_a)$ is accidentally supposed to have all the terms in the query. Thus, $D_1(T_a)$ has a high similarity value. On the other hand, the similarity calculation in TS_2 are based on all the terms T_a, T_b and T_c. The similarity between $D_2(T_a, T_b)$ and the query is lower even if $D_2(T_a, T_b)$ has T_a and T_b, because the document does not have T_c, which is in the query.

In order to make the similarity calculation more preferable, we have to properly treat unknown words in the query as the factor of discounting similarity in every LSI space, instead of just discarding them. As one of ways to do that, we propose the introduction of one new dimension into each LSI space to represent unknown words. This method adjust the similarity between documents and the query in terms of unknown words by expanding each LSI space and treat unknown words as one vector which is orthogonal with all other term vectors as follows.

Suppose that an LSI space is an n-dimensional space where a term is represented as a vector $(w_1, ..., w_n)$. We introduce one new dimension into the space and obtain an (n+1)-dimensional space so that each vector of existing (known) term is represented as $(w_1, \ldots, w_n, 0)$, in contrast, all unknown words in the query are represented as the vector $(0, \ldots, 0, 1)$. Since all of the (n+1)-th element of document documents are always zero(0), the following relations hold:

$$\begin{aligned}
\mathbf{D}' \cdot \mathbf{Q}' &= \mathbf{D} \cdot \mathbf{Q} \\
|\mathbf{D}'| &= |\mathbf{D}| \\
|\mathbf{Q}'| &= |\mathbf{Q} + \mathbf{Q_u}| \\
&= \sqrt{|\mathbf{Q}|^2 + |\mathbf{Q_u}|^2} \\
&= \sqrt{|\mathbf{Q}|^2 + (\sum_{T_i \in Q_u} tf(T_i, Q_u) idf(T_i))^2},
\end{aligned}$$

where

$\mathbf{D}$:	Document vector before the adjustment
$\mathbf{D}'$:	Document vector after the adjustment
$\mathbf{Q}$:	Query vector before the adjustment
$\mathbf{Q}'$:	Query vector after the adjustment
Q_u:	List of unknown words in the query Q
$\mathbf{Q_u}$:	Vector of Q_u.

The adjusted similarity $sim(\mathbf{D}', \mathbf{Q}')$ between the document vector $\mathbf{D}'$ and the query vector $\mathbf{Q}'$ in the new LSI space is given by the following formula.

$$\begin{aligned}
sim(\mathbf{D}', \mathbf{Q}') &= \frac{\mathbf{D}' \cdot \mathbf{Q}'}{|\mathbf{D}'||\mathbf{Q}'|} \\
&= \frac{\mathbf{D} \cdot \mathbf{Q}}{|\mathbf{D}| \sqrt{|\mathbf{Q}|^2 + (\sum_{T_i \in Q_u} tf(T_i, Q_u) idf(T_i))^2}}
\end{aligned}$$

As this formula shows, the similarity in the reconstructed space is discounted according to unknown words in the query.

6 Experimental Results

6.1 Extraction of Index Words from Documents

In our experiment, target documents are written in Japanese and English. Both simple words and compound words are used for indexing.

For Japanese documents, JUMAN 3.61(Kurohashi and Nagao, 1998) performs word segmentation and POS tagging. According to the POS information, nouns, adjectives, verbs, nominal modifier, adverbs, English words, KATAKANA words[3] are selected as simple index words.

For English documents, stemming algorithm by Frakes(Frakes, 1992) and the stop-word list described in Fox(Fox, 1992) are utilized.

As the uniform method to recognize compound words for any languages, we adopt the C-value(Frantzi and Ananiadou, 1996) based method with following steps. Firstly, the suffix array of word sequences in corpus are constructed in order to obtain the frequency of all word sequences. If the frequency of a word sequence is less than the threshold TH_f, it is discarded. Secondly, C value is calculated for remaining word sequences. We adopt as compound words the candidates which have C-value of TH_c or over.

In our experiment, NTCIR1 corpus are devided into eleven parts because of limitation of our program, and extract compound words from each part with the conditions $TH_f = 5$ and $TH_c = 5$. Constituents of compound words are also adopted as index words.

6.2 Experimental Results on Dividing LSI Space

There is the possibility that the segmented LSI would be less effective than the original monolithic LSI, because the segmented LSI does not refer to all of training corpus at once. On the ohter hand, there is the other possibility that the ambiguity in trans-lingual information would decrease and the performance would be improved, because the segmented LSI constructs LSI sub-spaces according to the document domains.

Thus, we conducted experiments of mate retrieval[4] in the following conditions, in which we can construct a monolithic LSI space with our computational environment.

We selected 6000 dual-language documents from the NTCIR1 corpus, and constructed two types of CL-LSI systems. First one has a monolithic LSI space, which is made from the whole of the document set. Second one has three LSI subspaces. As for the system, the document set was divided into three subsets according to area names and each LSI sub-space was made from each subset. In both of these systems, the dimension of term vectors are reduced to about 150 by SVDPACKC(Berry et al., 1993). We also selected the other 3000 dual-language documents for the evaluation by mate retrieval.

The results are shown in Table 1.

Table 1: Plural Spaces v.s. Monolithic Space

	Rank 1(%)	Within Rank 3(%)
Monolithic Space	58.2	75.7
Plural Spaces	47.8	63.9
Plural Spaces with Adjustment	59.4	78.2

6.3 Experimental Results of NTCIR2 J-E and E-J tasks

As a large-scale experiment, we participated in the Japanese-English CLIR task of NTCIR2(NTCIR Project, 2000). Since the evaluation of NTCIR2 is a large and practical scale, we can evaluate our scheme in the situation akin to real situations. We can use 380 thousand of summaries as a trainig corpus, and the document set to be retrieved contains 700 thousand of summaries. Those documents are written in Japanese and/or English.

In our experiment, as the training corpus, we extract all of Japanese-English pairs (about 180 thousand pairs) of summaries from the NTCIR1 corpora. Those summaries come from technical papers of 57 scientific societies. The corpus is divided into sub corpora by the algorithm in Section 4.1 as follows. We select six societies as major area groups. Then other area groups are merged to one of major area groups. According to our computational resources, each group of the four largest area groups is divided into two sub groups, and we finally obtained ten area groups. Each area group has from 14 to 26 thousand pairs of documents and from 78 to 115 thousand different terms. Total number of different terms in the corpus is about 380 thousand. For each area group, we obtain a set of term vectors (i.e. an LSI space) by the method described in section 2. The dimension of each LSI space is about 450 (from 430 to 463).

In the NTCIR2 test set, there are 49 topics of retrieval written in both Japanese and English. Although each topic has several fields, we use DESCRIPTION('DESC') field and DESCRIPTION+NARRATIVE('DESC-NAR') fields as

[3] KATAKANA words usually correspond to foreign words in Japanese.

[4] The mate retrieval is one of evaluation methods for CLIR. The one language part of each dual-language document is submitted as a query. Then, we examine the retrieval rank of its 'mate', namely, the paired document of it. If the average rank of retrieval is high, the method can be regarded as effective.

short query and long query, respectively. Our experimental results of NTCIR2 J-E (Japanese queries and English documents) and E-J (English queries and Japanese documents) tasks are shown in Table 2. The label 'with Adjust.' shows that it is the results with the adjustment of unknown words proposed in Section 5.

Table 2: Experimental Results of all Topics

	Average precision	R-Precision
J-E-DESC	0.0533	0.0635
with Adjust.	0.0666	0.0786
Gain by Adjust.	24.9 %	23.8 %
J-E-DESC-NAR	0.0868	0.1031
with Adjust.	0.0940	0.1096
Gain	8.3 %	6.3 %
E-J-DESC.	0.0512	0.0705
with Adjust.	0.0610	0.0839
Gain	19.1 %	19.0 %
E-J-DESC-NAR	0.0609	0.0876
with Adjust.	0.0736	0.1018
Gain	20.9 %	16.2 %

Since LSI is sensitive to unknow words in queries, we also examine the result of topics without unknown words as shown in Table 3. All keywords extracted from those topics can be found in dual-language documents used for building LSI spaces.

Table 3: Experimental Result of Topics with no Unknown Words

	n_q	Average precision	R-precision
J-E-DESC	43	0.0600	0.0704
with Adjust.	43	0.0743	0.0870
Gain by Adjust.		23.8 %	23.6 %
J-E-DESC-NAR	31	0.1032	0.1206
with Adjust.	31	0.1094	0.1307
Gain		6.0 %	8.4 %
E-J-DESC	43	0.0579	0.0786
with Adjust.	43	0.0692	0.0942
Gain		19.5 %	19.8 %
E-J-DESC-NAR	39	0.0738	0.1025
with Adjust.	39	0.0872	0.1187
Gain		18.2 %	15.8 %

n_q: Number of queries without unknown words.

7 Discussion

As the first experiment (Table 1) shows, the segmented LSI is less effective than the original LSI. However, by introducing the adjustment of unknown words, the effectiveness of segmented LSI is considerably improved. The revised segmented LSI has almost same or a little bit higher effectiveness than the original LSI. Those two results was expected in advance.

Next, let us examine the second experiment. In the viewpoint of absolute effectiveness of information retrieval, we have to say that our system, which is only based on a set of dual-language documents, is less effective than other systems which would be based on translation dictionaries. The best result in NTCIR2 participating systems is above 0.3 in the average precision, while our method achieves only about 0.1. However, we confirm that we can construct a large-scale CLIR system without dictionaries, if we have a enough number of dual-language documents. Moreover, there is plenty of room for improvement, if we introduce some (pseudo) relevance feedback mechanisms, which IR systems usually introduce as a well-worn device.

We also confirm that our adjusting method for unknown words is very effective. For example, the average precision of 'J-E-DESC' in Table 3 is improved in 24.9 %. The average precision of 'J-E-DESC-NAR' also rises in 8.3 %. The precision of the retrieval with a query made from the DESCRIPTION field only is improved more than the case that DESCRIPTION+NARRATIVE fields are used as a query. The reason is that shorter queries were relatively more affected by unknown words.

From the comparison between Table 2 and 3, we can find that totally unknown words still affect the performance of retrieval.

There are another participated system of NTCIR2 by Jian et al.(Jiang and Littman, 2001), which is based on an corpus-based approach like ours. They introduce a method called 'Approximate Dimension Equalization', which achieves the effect of LSI with smaller number of singular vector calculation. In the NTCIR2 evaluation, they report that the average precisions of J-E and E-J tasks are 0.0724 and 0.0829, respectively[5]. The average of them is 0.0777. The result is almost similar to ours, that is, 0.0838(the average of J-E and E-J for the field DESCRIPTION+NARRATIVE) and 0.0638(for the field DESCRIPTION).

One of the possible reasons why both of these methods, which are fully depends on training corpora, do not have higher effectiveness would be that the bilingual copus extracted from NTCIR1 corpus does not match with the documents in NTCIR2. We have not examined whether it is true or not, it shows the limitation of schemes which fully depends on training corpora.

8 Concluding Remarks

In this paper, we studied the CL-LSI method where a set of dual-language documents is only required to construct translation information for information retrieval. We proposed a way to apply it to a large set of dual-language documents by dividing the set into several subsets and constructing plural LSI spaces. We also study the de-

[5]We do not have information about what field is used as query.

cline in accuracy of retrieval, which is caused by difference in vocabularies of the LSI spaces. We showed that our adjustment for unknown words is effective to solve the problem.

In the viewpoint of absolute effectiveness of retrieval, we have to conclude that our system is less effective than other systems which is based on translation dictionaries. However, we reconfirm that we can construct a large-scale CLIR system without dictionaries, if we have a enough number of dual-language documents.

The following problems will be parts of our future works.

- Confirmation of improvement of accuracy by introducing plural LSI spaces.

  The experimental result of mate retrieval shows that the division of word spaces is effective to improve the precision of retrieval. However, it is not obvious how effective it is in real retrieval situations like NTCIR2. Additional experiments are needed to confirm the effectiveness.

- Estimating vector of unknown words.

  In the original LSI scheme, a method to estimating vector of unknown words from new documents is proposed. In the method, the vector of each unknown word is made by sum up the vectors of documents in which the unknown word appears. It would be possible to introduce the estimation to our method.

- Combination with (pseud) relevance feedbacks.

References

Michael Berry, Theresa Do, Gavin O'Brien, Vijay Krishna, and Sowmini Varadhan, 1993. *SVDPACKC (Version 1.0) User's Guide.* Computer Science Department, University Tennessee.

Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. 1997. Translingual information retrieval: A comparative evaluation. In *Proceedings of International Joint Conference on Artificial Intelligence '97 IJCAI '97.*

Scott Deerwester, Susan T. Dumais, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407.

Susan T. Dumais, Thomas K. Landauer, and Michael L. Littman. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In *SIGIR '96 – Workshop on Cross-Linguistic Information Retrieval*, pages 16–23.

Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, March.

Christopher Fox. 1992. Lexical analysis and stoplists. In William B. Frakes and Ricardo Baeza-Yates, editors, *Information Retrieval – Data Structure & Algorithms*, chapter 7, pages 102–130. Prentice Hall PTR.

William B. Frakes. 1992. Stemming algorithms. In William B. Frakes and Ricardo Baeza-Yates, editors, *Information Retrieval – Data Structure & Algorithms*, chapter 8, pages 131–160. Prentice Hall PTR.

K. Frantzi and S. Ananiadou. 1996. Extracting nested collocations. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96)*, pages 41–46, August.

David A. Hull and Gregory Grefenstette. 1996. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of SIGIR '96: 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–57.

Fan Jiang and Michael L. Littman. 2001. Approximate dimension reduction at ntcir. In *Proceedings of NTCIR Workshop 2 Meeting*, pages 5–179–5–74, 3.

Genichiro Kikui. 2000. Retrieving Documents Across Language-Barriers. *Journal of Japanese Society for Artificial Intelligence*, 15(4):550–558, July. (in Japanese).

Tadao Kurohashi and Makoto Nagao, 1998. *Japanese Morphological Analysis System JUMAN version 3.61 Manual.* Kyoto University. (in Japanese).

Sonia Leach. Singular valued decomposition — a primer. Department of Computer Science, Brown University.

NTCIR Project. 2000. NTCIR (NII-NACSIS test collection for IR systems) project web page. http://research.nii.ac.jp/ ntcadm/index-en.html.

Douglas W. Oard and Bonnie J. Dorr. 1996. A survey on multilingual text retrieval. Technical Report UMIACS-TR-96-19 CS-TR-3615, University of Maryland, April.