Cross Language Information Retrieval based on Comparable Corpora

S. Nakazawa

C&C Media Research Laboratories
NEC Corporation
nakazawa@ccm.cl.nec.co.jp

T. Ochiai

Open Technology Systems Division
NEC Information Systems, Ltd.
ochiai@ksp.nis.nec.co.jp

K. Satoh **A. Okumura**

C&C Media Research Laboratories
NEC Corporation
{satoh,okumura}@ccm.cl.nec.co.jp

Abstract

This paper describes our method for the Cross Language Information Retrieval (CLIR) task of NTCIR. The query terms in our method are expanded in the source language first, and then the expanded terms are translated stepwise. Our method utilizes term co-occurrence frequency in comparable corpora and a bilingual dictionary.

Keywords : Cross Language Information Retrieval, Comparable Corpora, Query Term Expansion, Stepwise GDMAX, co-occurrence frequency

1 Introduction

The explosive growth of the World Wide Web has increased the necessity for multi-lingual information retrieval systems, especially CLIR, which accepts a query in one language and retrieves relevant documents in other languages. For example, CLIR makes it possible to retrieve English documents by using Japanese retrieval requests(queries). And CLIR is considerably more complex than mono-lingual information retrieval because a method for translating queries or documents must be developed before one can use mono-lingual information retrieval (IR) [1].

Most CLIR is done by query translation and information retrieval. However, CLIR is less precise than IR because of ambiguities in translation of query terms,especially in Japanese and English CLIR [2].

The GDMAX method, which can resolve this ambiguity problem found in CLIR, was proposed [3]. In the GDMAX method, term co-occurrence frequency data are calculated in both the source language corpora and target language corpora. Query terms are translated based on these co-occurrence data and a bilingual dictionary.

The GDMAX method has some advantages : it can translate one query term into several appropriate target terms in CLIR and, it only needs easily available linguistic resources, etc. However, it requires a great amount of calculation power. Therefore we have proposed the improved query term translation method for CLIR.

2 Selection of Target query terms

As listed in Table 1, a source query term j_i generally has target-term ambiguities, e_{i1}, ..., e_{ik} ,..., e_{ir} , in a bilingual dictionary. An appropriate set of target query terms should be selected from the target-term ambiguities of each source-query term.

Table 1: Selection of Target query terms

Source query terms	Target query terms
j_1	e_{11} e_{12} ... e_{1p}
j_2	e_{21} e_{22} ... e_{2q}
...	...
j_i	e_{i1} ... e_{ik} ... e_{ir}
...	...
j_n	e_{n1} e_{n2} ... e_{nm}

There are three kinds of query term translation methods. The first method is based on corpora which uses parallel corpora or comparable corpora. The second method is based on bilingual dictionaries (including the machine translation method). And the third method is the hybrid method which synthesizes the above two methods [1].

Although parallel corpora are more precise, they are not common in general language, tending instead to be restricted to specialized domains. Although comparable corpora are easier to collect, they are likely to incorporate low-relevance terms into queries.

A dictionary alone is not flexible in selecting terms according to domains. But a bilingual dictionary and

comparable corpora used together can compliment each other.

3 GDMAX method

The GDMAX method has a hybrid architecture; that is, it uses comparable corpora to select a set of target query terms from all ambiguities described in a bilingual dictionary. GDMAX searches target translation terms by considering all possible term pairs in order to give a logical ranking order to all the combinations of target query terms.

At the end of this section, we will briefly explain the algorithm of the GDMAX method.

In the GDMAX method, the term co-occurrence frequency data is used to generate term co-occurrence frequency vectors (TCFVs) from a source query and all possible translation queries. A source query TCFV and the target query TCFVs are respectively based on term co-occurrence frequency data of source corpora and those of the comparable target corpora.

For example, Japanese queries consisting of n terms generate a list of TCFVs, $\mathbf{F}_{jap}$, as

$$\mathbf{F}_{jap} = (\mathbf{f}_j^1, \mathbf{f}_j^2, .., \mathbf{f}_j^n) \quad (1)$$

where $\mathbf{f}_j^p$ is a ${}_nC_p$ dimension TCFV, which is composed of p-term co-occurrence frequencies in one document. In other words, instead of terms in a regular vector-space model, the GDMAX method uses term co-occurrence frequencies to represent a query.

The following equation shows that $\mathbf{f}_j^2$ consists of the co-occurrence frequencies of two arbitrary terms in one document.

$$\mathbf{f}_j^2 = (f(j_1, j_2), f(j_1, j_3), ...f(j_{n-1}, j_n)) \quad (2)$$

Here, $f(j_i, j_j)$ is a normalized value of the co-occurrence frequency of terms j_i and j_j in the Japanese corpus.

In the same way as Japanese TCFVs, a list of TCFVs generated from an English translation query, $\mathbf{F}_{eng}$, is represented as

$$\mathbf{F}_{eng} = (\mathbf{f}_e^1, \mathbf{f}_e^2, .., \mathbf{f}_e^n) \quad (3)$$

For example, given the translation ambiguities of query terms listed in Table 1, there are $p*q*r*..*m$ alternatives for $\mathbf{F}_{eng}$.

Similarity $\mathbf{Sim}(\mathbf{F}_{jap}, \mathbf{F}_{eng})$ is a function of $\mathbf{Sim}(\mathbf{f}_j^1, \mathbf{f}_e^1)$,$\mathbf{Sim}(\mathbf{f}_j^2, \mathbf{f}_e^2)$,... and $\mathbf{Sim}(\mathbf{f}_j^n, \mathbf{f}_e^n)$. Here, similarity $\mathbf{Sim}(\mathbf{f}_j^p, \mathbf{f}_e^p)$ is defined as

$$\mathbf{Sim}(\mathbf{f}_j^p, \mathbf{f}_e^p) = \frac{(\mathbf{f}_j^p, \mathbf{f}_e^p)}{|\mathbf{f}_j^p||\mathbf{f}_e^p|} \quad (4)$$

In practice, co-occurrence frequencies of more than two terms are negligible in comparison with other frequencies because of data sparseness. Although single-term frequency is huge, it should be suppressed because Japanese terms have a completely different ambiguity set from that of English equivalent terms. For example, the Japanese term "米" means "rice" and "USA". The huge frequency of "米" in the Japanese corpus does not always mean that "rice" frequently appears in the comparable English corpus. The co-occurrence frequencies of two terms are therefore used in this model. That is, $\mathbf{F}_{jap}$ and $\mathbf{F}_{eng}$ are matched by using $\mathbf{f}_j^2$ and $\mathbf{f}_e^2$. The matching is done by calculating the inner product and the cosine between $\mathbf{F}_{jap}$ and the alternatives for $\mathbf{F}_{eng}$, as shown in the following:

$$\begin{aligned} \mathbf{Sim}(\mathbf{F}_{jap}, \mathbf{F}_{eng}) &= \mathbf{Sim}(\mathbf{f}_j^2, \mathbf{f}_e^2) \\ &= \frac{(\mathbf{f}_j^2, \mathbf{f}_e^2)}{|\mathbf{f}_j^2||\mathbf{f}_e^2|}. \end{aligned} \quad (5)$$

The alternatives can be ordered according to the vector similarity ranking.

For example, the Japanese query terms j_1, j_2, and j_3 have three values, $f(j_1, j_2)$, $f(j_1, j_3)$, and $f(j_2, j_3)$ for two-term frequencies. They are illustrated in the form of a triangle in the Japanese corpus space as shown in Figure 1. The GDMAX method searches the comparable English corpus for triangles of e_{1i}, e_{2j}, e_{3k} similar to the triangle of j_1, j_2, j_3. English query terms are a set of English TCFVs whose similarity is more than a given threshold.

3.1 Advantages and Problems

The GDMAX method has the following advantages.

The first advantage is that GDMAX can change the similarity threshold according to purpose of CLIR, because GDMAX ranks all possible target query term combinations.

The second advantage is that GDMAX only needs a bilingual dictionary and comparable corpora, which are more accessible than parallel corpora. And term co-occurrence data can be counted automatically, so

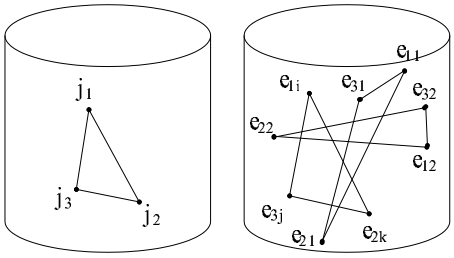

Figure 1: TCFV Similarity in comparable corpora

a person dosn't need to decide the translation rules which he/she must decide in a machine translation method.

However, the following problems exist:

The first problem is that GDMAX requires a great amount of calculating time. GDMAX calculates Eqn.5 for all possible target term combinations, and ranks them. Therefore GDMAX must calculate the similarity measure as many times as the number of all possible target term combinations. For example, GDMAX calculates $p \times q \times \cdots \times r \times \cdots m$ times in the case of Tab.1. Therefore, an increasing number of query terms causes the calculation time to increase exponentially. This means that we can not use GDMAX correctly to translate more than relatively few query terms.

The second problem is data sparseness. Co-occurrence frequency data are used in Eqn.1 and Eqn.3. If one of these vector value become 0, the denominator of Eqn.5 also become 0, and we can't use the GDMAX method in this case. Even if no vector become 0, the GDMAX become inaccurate when co-occurrence frequency data are less.

The first problem becomes important when there are a lot of query terms and possible target term combinations. The second problem becomes important when there are few query terms.

4 Proposal method

In this paper, we propose a new CLIR method, which has the same advantages GDMAX, and resolves the two problems(large calculation time and that were we explained at subsection 3.1. Basically our new CLIR method consists of query term expansion and stepwise GDMAX.

Query term expansion resolves the inaccuracy with co-occurrence frequency data. By using this expansion, GDMAX can use more co-occurrence frequency data.

Stepwise GDMAX resolves the calculation time problem. GDMAX calculation time depends on the number of all possible target term combinations. We can reduce this number by translating query terms stepwise.

4.1 Query Term Expansion

Query term expansion is the technique for Information Retrieval that expands input query terms and include their synonyms as well. This is effective for a retrieval leak. In general, query term expansion is used for mono-lingual IR.

When query term expansion is used for CLIR, there are two possible methods: The first method expands the terms in the source language, and then both the original and expanded terms are translated into the target language. The second method translates the original terms first, and then expands them in the target language. We use the first method for the following reasons.

1. In the case where there are not too many input source query terms,
   (a) GDMAX does not obtain enough co-occurrence data for translation clue.
   (b) So add related terms by query term expansion
   $\Rightarrow$ GDMAX can use more co-occurrence data.
2. In case there are many input source query terms.
   (a) Do the query term expansion anyway.
   (b) When there are too many possible target term combinations, we translate original query terms preferentially, then translate the expanded terms. Stepwise GDMAX can evade the calculation time problem this way.
3. In any case, added synonyms make IR more accurate.

Our query term expansion algorithm is shown as follows.

Algorithm of Query Term Expansion

1. Extract query terms from the input query by morphological processing. Then remove stop words from query terms.

2. For any query term j_i,

$$\frac{f(j_i, j_a)}{f(j_i) \cdot f(j_a)} \geq TETH1 \quad (6)$$

seek every term j_a which satisfies the above condition. Those terms make the expansion candidate list for the original term j_i. Here $f(j_i, j_a)$ is the co-occurrence frequency of term j_i and term j_a. $f(j_i)$ and $f(j_a)$ means the occurrence frequency of term j_i and term j_a. $TETH1$ is the adequate threshold. Expansion candidate lists are independent of queries, so we calculate those lists beforehand.

3. Here $j_1 \cdots j_n$ are original query terms. Seek all the terms that satisfy the condition below from all expansion candidate lists.

$$\sum_{i=1}^{n} \frac{f(j_i, j_a)}{f(j_i) \cdot f(j_a)} \geq TETH2 \quad (7)$$

We adopt those as final expanded terms. $TETH2$ in the above condition is the adequate threshold. In stead of using this threshold, we may adopt the highest l terms of the above condition's left value.

Our query term expansion has two phases. In the first phase, one original query term j_i is focused on at a time. All the terms which co-occur with j_i are included in the expansion candidate list for the original term j_i. Then each expansion candidate list for all original terms are calculated one by one. In the second phase, the summation of co-occurrence frequency with all original terms are calculated for each candidate term. The candidate terms which have a large summation are selected as expanded query terms.

In the general query term expansion method based on co-occurrence frequency, the terms which co-occur with one original term are added as expanded query terms. In our method, we use the first phase as a filter, and in the second phase we select terms which co-occur equally with all original query terms.

For example, one term a co-occurs with one original query term j_1 when j_1 is used as the meaning s_1, but in this time query, j_1 is used for meaning s_2. Our method prevents the case where term a is selected as an expanded term by mistake. In other words, few terms which have a very different meaning from the given query are selected as expanded query terms.

4.2 Stepwise GDMAX

After we have finished the query term expansion, we seek the target query term lists as Table 1 for each the original and expanded term by looking up a bilingual dictionary.

Increasing the number of all possible target query term combinations causes the calculation time problem we explained at subsection 3.1. Therefore the basic idea of the stepwise GDMAX we will describe at this subsection is to devide the translation processes and reduce all possible target query term combinations we must handle at a time.

Our stepwise GDMAX algorithm is shown as follows.

Algorithm of Stepwise GDMAX

1. To reduce the number of target terms in target query term lists, we make virtual target terms from synonyms in a target query term list. In other words, we consider a few synonyms in a target query term list as one virtual target term. For example we consider the term "Corp" and the term "Corporation" as one virtual term. We use the average of each co-occurrence data as virtual term co-occurrence data.

2. We decide the source query term group division, and the schedule of GDMAX, based on the following rules.

   (a) We consider the source query terms which have a large number of target query terms as very ambiguous terms. We do GDMAX for this kind of terms at last.

   (b) We prefer the original query terms to the expanded terms. One expanded term are translated at a time with the original terms which have already been translated.

   (c) We group source terms which co-occur with one another, based on the left-value of Eqn.6.

3. We apply the GDMAX method to one group at a time according to the schedule.

4. When we have finished the stepwise GDMAX, we output the target query term sets to an IR engine.

5 Experiments

We made experiments for the NACSIS Test Collection CLIR task[5].

We considered the NTCIR corpora NTC1-J0 and NTC1-E0 as Japanese-English comparable corpora.

And we counted term co-occurrence frequency data from the $< TITL >$ parts, the $< ABST >$ parts, and the $< KYWD >$ parts of each corpus independently. We counted the following way: when term a occurs more than once in one record, and also term b occurs more than once in the same record, we count just one co-occurrence with term a and b.

The experiment method is basically the same as we expained at the section4. We will describe only experiment conditions at the following two subsections.

5.1 Bilingual Dictionaries

Bilingual dictionaries have great effect on CLIR results. We must choose adequate dictionaries according to IR object domain.

This time we used the basic and technical dictionaries we had for the machine translation. And to these dictionaires, we added keyword pairs which were used in $< KYWD >$ parts of NTCIR corpora.

The total Japanese entry number is 464205.

And we used the same dictionaries in morphological process of the queries. We adopted only noun terms, 'Sa-hen' verb terms, and unknown terms from the queries. Then we removed the terms which were on the stop-word list (about 200 words).

5.2 Other experiment conditions

When we expanded query terms, we set the threshold in Eqn.7 $TETH2 = 3 \times 10^{-5}$. We ranked all the terms which satisfied this equation, and adopted the terms twice as many as the original terms.

When we made target query term lists, we added the source term itself which was written in alphabet to its target query term list.

The threshold of Eqn.5 was 0.85. We used all co-occurrence frequency data from $< ABST >$ parts, $< TITL >$ parts, and $< KYWD >$ parts of corpora, but gave -15 bias to the co-occurrence frequency of $< ABST >$ parts.

We used the IR engine SMART[4] which is based on the word vector space model as an English IR engine. We didn't adjust word weights when we made SMART IR indices. But when we gave target query terms to the SMART, we weighed the target terms of the original terms three times as much as the target terms of the expanded terms.

5.3 Experimental results

Figure 2 shows the Precision-Recall curves of CLIR results (averaged over queries). Table 3 shows the average precision of each method. Each method is explained in Table 2.

We used only NTCIR official 39 queries. But in "Expanded and translated by a person" method, we only used 10 queries. Because the person must know the query domains well.

Each method produced English query terms and inputed them to the SMART.

Our method took 6 minutes and 17 seconds for the query term expansion and the stepwise GDMAX of 53 queries on Sun UltraSPARC-II(296MHz, 524Mbyte). On the other hand, over 8 hours calculation was not enough for the orginal GDMAX method. Therefore we stopped the GDMAX calculation.

6 Discussions of the results

IR results of all methods including the method by a person are totally low. We didn't adjust the term weight of SMART indices. This may be the reason of low results. Automatic adjustment according to the queries is one of the next problems which we will tackle.

The results of our method are better than the machine translation method by the average precison measure. "Only stepwise GDMAX" method is the worst among 4 methods. This is because term co-occurrence frequency vectors became 0 for some queries. As we explained at subsection 3.1, we can't use GDMAX in these cases, so all possible target terms were given to SMART. This result shows the effect of the query term expansion in the source language.

We couldn't measure the calculation time of the normal GDMAX method. But it took at least 8 hours. Our method took 6 min. 17 sec. for the same problem. There is a large difference between the normal GDMAX and our method from the viewpoint of calculation time.

As we mentioned before, automatic adjustment of SMART term weights is one of the next tasks. Automatic decision of thresholds that our method uses is another task. Also we will research another query term expansion method.

Table 2: Methods to produce English query terms

Proposal method	results by our method (the query term expansion and the stepwise GDMAX)
Only stepwise GDMAX	results by our method but without the query term expansion
CROSS	Machince Translation method. (NEC Translation Adaptor II with technical dictionaries of information, electricity, and science)
Expanded and translated by a person	the person who knows the query domains well translated the queries and expanded terms.

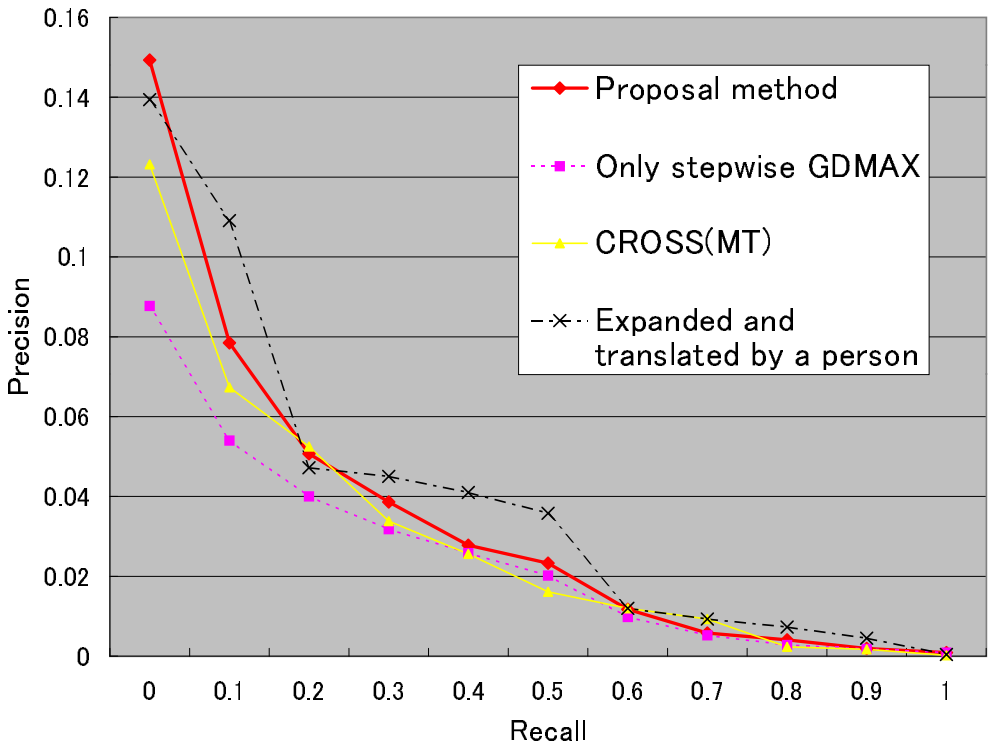

Figure 2: Precision-Recall curves

Table 3: Average Precision(averaged over queries)

Proposal method	0.0298
Only stepwise GDMAX	0.0210
CROSS(MT)	0.0246
Expanded and translated by a person	0.0351

References

[1] David Hull : "Using Structured Queries for Disambiguation in Cross-Language Information Retrieval", Technical Report of 1997 AAAI Spring Symposium on Cross-Language Text and Speech Retrieval (1997).

[2] K. Yamabana, K. Muraki, S. Doi, and S. Kamei : "A Language Conversion Front-End for Cross-Language Information Retrieval", Workshop on Cross-Linguistic Information Retrieval SIGIR'96 (1996).

[3] A. Okumura, K. Ishikawa, and K. Satoh : "Japanese-English Cross Language Information Retrieval based on Comparable Corpora and Bilingual Dictionary", Journal of Natural Language Processing, Vol.5, No.4 (1998.10).

[4] Gerard Salton : "The SMART Retrieval System : Experiments in Automatic Document Processing", Prentice-Hall (1971).

[5] NACSIS : "NACSIS Test Collection for IR Systems Project", http://www.rd.nacsis.ac.jp/ ntcadm/index-en.html