Query Term Disambiguation for Web Cross-Language Information Retrieval using a Search Engine

Akira Maeda, Fatiha Sadat, Masatoshi Yoshikawa, and Shunsuke Uemura

Graduate School of Information Science
Nara Institute of Science and Technology (NAIST), Japan
Email: aki-mae@is.aist-nara.ac.jp, fatia-s@is.aist-nara.ac.jp, yosikawa@is.aist-nara.ac.jp, uemura@is.aist-nara.ac.jp

Abstract

With the worldwide growth of the Internet, research on Cross-Language Information Retrieval (CLIR) is being paid much attention. Existing CLIR approaches based on query translation require parallel corpora or comparable corpora for the disambiguation of translated query terms. However, those natural language resources are not readily available. In this paper, we propose a disambiguation method for dictionary-based query translation that is independent of the availability of such scarce language resources, while achieving adequate retrieval effectiveness by utilizing Web documents as a corpus and using co-occurrence information between terms within that corpus. In the experiments, our method achieved 97% of manual translation case in terms of the average precision.

Keywords: cross-language information retrieval; mutual information; search engine; WWW

1 Introduction

With the increasing popularity of the Internet in various parts of the world, languages used for Web documents are expanded from English to others. However, there are many unsolved problems in order to realize a retrieval system that can handle such multilingual documents in unified manner. For instance, although character coding systems used for Web documents vary according to the languages, many Web documents do not have meta information of the coding system of the document itself. Also, fonts and input methods, which are necessary for displaying and inputting characters, are not always installed in the user's terminal for particular languages. Depending on those problems, some solutions were already developed, such as an automatic identification of coding systems of documents[1] and a multilingual browser for Web documents, which does not require fonts in user's terminal[2].

Some Web search engines such as AltaVista and Lycos can handle multiple languages in addition to English and can specify the target language of the documents to be retrieved.

Proceedings of the 5th International Workshop Information Retrieval with Asian Languages

In the same time, many search engines exist that handle Web documents written in a particular language other than English. However, these search engines are essentially a collection of monolingual search engines from the user's perspective. Nevertheless, there might be some cases where the user wishes to retrieve documents in unfamiliar languages. Needs for retrieving such information must be large. For example, when Japanese is used as a query language, target collection will represent only a very small portion of the whole Web documents, which consist of several hundreds millions of pages. Also, there might be cases, depending on the user's demand, where information written in a language other than the user's native language is rich. For example, for the economic trend of a particular country, there should be extremely rich information in the language related to that country.

To fulfill such needs, researches on Cross-Language Information Retrieval (CLIR), a technique to retrieve documents written in a certain language using a query written in another language, have been active in recent years. Of course, an obvious solution is to translate all Web documents into the query language in advance. However, considering the enormous amount of Web documents, this approach is unrealistic. Therefore, it is feasible to translate the retrieved documents into the specified language. Recently, relatively inexpensive machine translation softwares are becoming more available. Some of them can translate and display a Web document on a Web browser on-the-fly. Therefore, some users cannot use languages other than their own native language in this case, increasing solutions in CLIR can be considered as useful.

Therefore, there is a problem in the query formulation in the target language. In order to satisfy such needs on usual monolingual retrieval system, the user has to manually translate the query by using a dictionary. This process not only imposes a burden to the user but also might choose incorrect translations for the query, especially for languages that are unfamiliar to the user. Accordingly, the approach in which the user formulates a query in his/her native language and the system translates it, should be desirable. One of the major technical problems to be solved in CLIR concerns the translation of short queries of one or few words, appropriately. Possible translation-candidates might be numerous in such cases and resolving such ambiguities becomes a hard task.

In this paper, we propose a novel approach for CLIR system targeting Web documents, which uses a natural language resource that is extracted from a Web search engine as a corpus, and resolves the ambiguities caused by the dictionary-based query translation approach, by using a co-occurrence information. We have evaluated the effectiveness of this method by experiments. By using this method, we do not have to worry about obtaining expensive language resources, which is one of the inconveniences of existing CLIR approaches. Easy extension to other languages, as well as the achievement of a reasonable retrieval effectiveness are among the advantages of our approach. We also conducted a comparative evaluation of four co-occurrence measures; mutual information, modified Dice coefficient, log likelihood ratio, and χ^2 test.

2 Related Work

Approaches to CLIR can be classified into three categories; document translation, query translation, and the use of inter-lingual representation. The approach based on translation of target documents has the advantage of utilizing existing machine translation systems, in which more context information can be used for disambiguation. Thus, in general, it achieves a better retrieval effectiveness than those based on query translation. However, since it is impractical to translate a huge document collection beforehand and it is difficult to extend this method to new languages, this approach is not suitable for multilingual, large-scale, and frequently-updated collection of WWW. The second approach transfers both documents and queries into an inter-lingual representation, such as bilingual thesaurus classes or a language-independent vector space. The latter approach requires a training phase using a bilingual (parallel or comparable) corpus as a training data.

The major problem in the approach based on the translation and the disambiguation of queries is that the queries submitted from ordinary users of Web search engines tend to be very short (approximately two words on average[3]) and usually consist of just an enumeration of keywords (i.e. no context). However, this approach has an advantage that the translated queries can simply be fed into existing monolingual search engines. In this approach, a source language query is first translated into the target language using a bilingual dictionary, and then the translated query is disambiguated. Our method falls into this category. One of the crucial problems of dictionary-based translation is the lack of headwords (e.g. compound words, coined words, loan words, etc.). Fujii et al.[4] proposes methods for the translation of compound words and transliteration of phonograms (i.e. katakana in Japanese) using bigram statistics.

For the disambiguation method, various approaches have been proposed[5, 6], such as using the first term listed in the dictionary, using relevance feedback, and using a parallel or a comparable corpus. However, such bilingual corpora are usually not readily available. Nie et al.[7] proposes a method to automatically gather parallel texts from the Web and use them for query term selection. However, for language pairs other than English-French in their case, the amount of parallel documents on the Web might not always be enough. Therefore, query term disambiguation method that does not depend on expensive language resources such as parallel corpus, is of practical use. For such a type of methods, *mutual information*, which is calculated from co-occurrence frequency of terms in a monolingual corpus, has been employed in several researches[8, 9, 10].

Lin et al.[9] uses mutual information for the disambiguation of translated queries in Japanese-English CLIR task. However, their method selects the only translation-candidate that has the highest score for actual retrieval. In this case, there is a possibility of selecting inappropriate translation. Jang et al.[10] uses mutual information for both disambiguation and query term weighting in Korean-English CLIR task. In their experiments, the method achieved 85% of monolingual retrieval case. According to Lin et al. and Jang et al. experiments, the window size of the co-occurrence was set to 3 and 6 words, respectively. In proportion the narrower the window size, the less effect of unrelated terms. However, term pairs that do not co-occur will increase even if the corpus is relatively large. Furthermore, mutual information has an undesired characteristic, which is the assignment of unexpectedly high values to rarely occurred terms[11].

Although our method[8, 12] also uses mutual information. In order to avoid the possibility of selecting only inappropriate translations, we take all translation-candidates that exceed a certain threshold value. In addition, in order to avoid the undesired effect of rarely occurred terms, we set another threshold value for the minimum occurrence of a term in a corpus. Moreover, by utilizing a Web search engine, we make it possible to use it as a huge corpus of various domains for the disambiguation without collecting enormous amount of Web documents.

3 Query Translation

Figure 1 shows the flow of query translation for the proposed query term disambiguation method. A query in

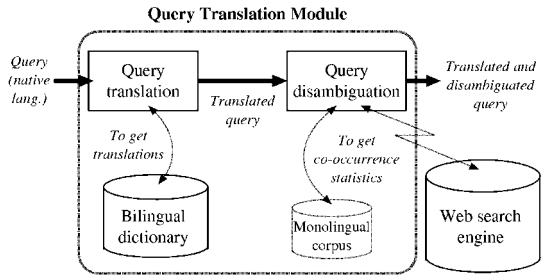

Figure 1: Flow of query translation.

user's native language is first translated into the target language, using a bilingual dictionary. The obtained translation-candidates are disambiguated, using term co-occurrence statistics and then passed to the search engine.

A query submitted by a user is first segmented into words using a morphological analyzer. Then, each word is translated into the target language using a machine-readable dictionary. In this phase, the longest matched term in the dictionary is used as the translation term. For example, an English query "digital library" can be segmented into "digital" and "library", but if the phrase "digital library" was found in the dictionary, the translation(s) of that phrase will be used instead. For the case of the longest match overlaps (e.g. "distributed network environment" matches both "distributed network" and "network environment"), both phrases will be used. Translation candidates obtained from the dictionary are then disambiguated using the method described in the next section, which is based on 2 or *n* words co-occurrence frequency information obtained from the target language corpus.

4 Query Term Disambiguation using a Web Search Engine

Since Web search engines gather an enormous volume of documents that cover extensive domains, they might be very useful as natural language resources. For instance, the number of retrieved documents by searching some terms combined by AND operators, can be regarded as a co-occurrence frequency of those terms in a Web document corpus. Ikeno et al.[13] investigates the possibility to apply it for selecting appropriate translated words for machine translation. We apply this method to the query disambiguation in CLIR.

4.1 Different Measures of Co-occurrence

Here we define several different measures of co-occurrence (we call them *co-occurrence tendency*). First, for two words w_1 and w_2 to calculate co-occurrence tendency, we define n_{ij} (i, j=1, 2) in a 2-by-2 table shown in Table 1. In the table, n_{11} indicates the number of times two words w_1 and w_2 co-occur within a text window, n_{12} indicates the number of times w_1 occurs, but w_2 does not occur within a text window, and so on.

Table 1: 2-by-2 table for calculating two words co-occurrence tendency.

	w_2	$\neg w_2$
w_1	n_{11}	n_{12}
$\neg w_1$	n_{21}	n_{22}

$n_{i\cdot}$, $n_{\cdot j}$, and N are defined as follows:

$$n_{i\cdot} = n_{i1} + n_{i2},\ n_{\cdot j} = n_{1j} + n_{2j},\ N = \sum_{i,j} n_{ij}$$

That is, $n_{i\cdot}$ indicates the number of times w_1 occurs (i=1) or does not occur (i=2) regardless of the occurrence of w_2, and N indicates the total number of co-occurrence windows in the corpus.

Generally, the window size of co-occurrence is a fixed number of words, but mainly for the limitation of utilizing search engines, we use one document as the window of co-occurrence. Actually, for example $w_1 \wedge \neg w_2$ (n_{12} in Table 1) can be obtained by submitting a query "w_1 AND NOT w_2" (in syntax of AltaVista). In this case, N is the total number of documents in the search engine. For example, for AltaVista, the total number of documents can be obtained by submitting a query of "*" symbol that matches arbitrary strings.

4.1.1 Mutual Information

Mutual Information is one of the metrics that can be used for calculating the significance of word co-occurrence associations[14]. MI can be applied to the words in the documents and can be used to calculate the correlation between those words. Two words co-occurrence tendency COT_{MI2} between words w_1 and w_2 in a corpus is defined as follows:

$$COT_{MI}(w_1, w_2) = \log_2 \frac{\frac{n_{11}}{N}}{\frac{n_{\cdot 1}}{N}\frac{n_{1\cdot}}{N}} \quad (1)$$

Note that we use one document as the window of co-occurrence instead of fixed number of words.

Usually, co-occurrences are measured between two words mainly because of the computational and the storage cost. However, when using a Web search engine, it is not necessary to calculate the frequencies for every pair of words in advance. We can use co-occurrence frequencies among any n words. n words co-occurrence tendency COT_{MIn} among words w_1, w_2...w_n is defined, as an extension of COT_{MI2}, as follows:

$$COT_{MIn}(w_1, w_2 \ldots w_n) = \frac{1}{n-1}\log_2 \frac{\frac{f(w_1, w_2 \ldots w_n)}{N}}{\frac{f(w_1)}{N}\frac{f(w_2)}{N}\cdots\frac{f(w_n)}{N}} \quad (2)$$

where N is the total number of documents in the search engine, $f(w)$ is the number of retrieved documents for the word w, and $f(w_1, w_2, \ldots w_n)$ is the number of retrieved documents for the words $w_1, w_2 \ldots w_n$ combined using AND operators. Note that MI is essentially a measure between two events, so this is an ad hoc extension only for the purpose of calculating n words co-occurrence tendency.

4.1.2 Modified Dice Coefficient

Dice coefficient is rather a heuristic measure and it is commonly used for calculating the word co-occurrence tendency. Kitamura et al.[15] proposes a modification of Dice coefficient, which improves the accuracy of co-occurrence tendency by adding a weight, based on co-occurrence frequency. In this paper, we call it *modified Dice coefficient*. Two words co-occurrence tendency COT_{DICE} between words w_1 and w_2 in a corpus, is defined as follows:

$$COT_{DICE}(w_1, w_2) = (\log_2 n_{11}) \frac{2n_{11}}{n_{1\cdot} + n_{\cdot 1}} \quad (3)$$

4.1.3 Log Likelihood Ratio

Likelihood ratio test is a kind of hypothesis testing. It can also be used for calculating *COT*, and it is said to be more accurate than MI, especially for rare occurred terms[16]. Two words co-occurrence tendency COT_{LLR} between words w_1 and w_2 in a corpus, is defined as follows:

$$COT_{LLR}(w_1, w_2) = 2\sum_{i,j} n_{ij} \left(\log_2 \frac{n_{ij}}{N} - \log_2 \frac{n_{i\cdot} n_{\cdot j}}{N} \right) + \Delta \quad (4)$$

4.1.4 Chi-Square Test

χ^2 test is also a hypothesis testing and is used for testing the dependence of two variables whose probabilities are approximately χ^2 distributed. Since χ^2 is a continuous distribution, some corrections are needed for small counts. Yate's correction is applied if any of n_{ij} is smaller than 5. Two words co-occurrence tendency COT_{CHI} between words w_1 and w_2 in a corpus is defined as follows:

$$COT_{CHI}(w_1, w_2) = \begin{cases} \dfrac{N(|n_{11}n_{22} - n_{12}n_{21}| - \frac{N}{2})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}, & \text{if } \min(n_{ij}) < 5, \\ \dfrac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}, & \text{otherwise} \end{cases} \quad (5)$$

4.2 Comparison of COT Measures

Table 2 shows the comparison of various measures for the co-occurrence tendency of Japanese words "大気 (air)" and "汚染 (pollution)". In this example, the pair "air" and "pollution" is the most appropriate translation. It is ranked top for COT_{DICE}, COT_{LLR}, COT_{CHI}, but third for COT_{MI2}.

Table 2: Comparison of various measures for co-occurrence tendency of Japanese words.

	COT_{MI2}	COT_{DICE}
1	consideration contamination	air pollution
2	consideration pollution	consideration pollution
3	air pollution	atmosphere pollution
	COT_{LLR}	COT_{CHI}
1	air pollution	air pollution
2	consideration pollution	consideration pollution
3	air contamination	air contamination

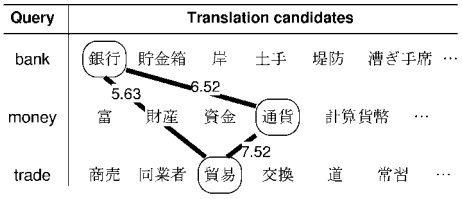

Figure 2: Example of a disambiguation using two words COT (COT_{MI2}).

4.3 Selection of Translations

When using two words co-occurrence tendency, the terms actually used for the query in target language are determined as follows:

1. Obtain the number of retrieved documents for each term in the query from the Web search engine,
2. Obtain numbers of retrieved documents for all possible combinations of each pair of translation-candidates, whose occurrence frequency for each term exceed the threshold value T_{freq}, from the Web search engine (using an AND operator),
3. Calculate the average of *COT*s for all possible combinations of the translation-candidate pairs,
4. The term sets whose *COT* exceed the threshold value T_{COT} are selected as the target language query.

Figure 2 shows an example of the disambiguation for the Japanese translation of an English query, which consists of three words "bank", "money", and "trade" using COT_{MI2} and the method above. In this case, the term set "銀行", "通貨", and "貿易" is given the highest average COT_{MI2}. In fact, these terms are the most appropriate translations for the source English words. Actually, all term sets which exceed T_{COT} are combined with OR operators, and used as a Japanese query.

We eliminate terms that rarely occur, in order to avoid the undesired phenomenon of MI described before. For example, in Figure 2, if we do not eliminate rarely occurred

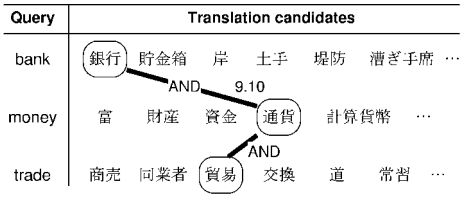

Figure 3: Example of a disambiguation using n words COT (COT_{MIn}).

terms, unrelated term sets "漕ぎ手席 (seat of an oarsman or a rower)", "富 (wealth or fortune)" and "常習 (habitual for customary)" is mistakenly given the highest average COT_{MI2} score.

On the other hand, when using n words co-occurrence tendency, the terms actually used for the target language query are determined as follows:

1. Obtain the number of retrieved documents for each term in the query from the Web search engine,
2. Obtain numbers of retrieved documents for all possible combinations of translation-candidates whose occurrence frequency for each term exceed the threshold value T_{freq}, from the Web search engine,
3. The term sets whose COT exceed the threshold value T_{COT} are selected as the target language query.

Figure 3 shows an example of a disambiguation using COT_{MIn} for the same query as the previous example. Also in this case, the same term set, "銀行", "通貨", and "貿易" is given the highest COT_{MIn}. All term sets which exceed T_{COT} are combined with OR operators, and are used as a Japanese query. This method may cost much time for querying the search engine, especially for queries with many possible translation-candidate pairs. However, it can be greatly reduced by submitting multiple requests for a query in parallel.

We can consider using a proximity operator instead of an AND operator. The proximity operator matches documents containing specified terms within a specific window of words, regarding or regardless of order. For example, AltaVista supports the proximity operator called "NEAR" which retrieves specified terms within 10 words regardless of order. In this case, to be exact, N is the total number of object windows, which cannot be calculated exactly using a search engine. However, since it does not affect the ranking of translation-candidates, and since the absolute value is not important in this case, we used the total number of documents for N in our experiments.

5 Evaluation

We have conducted some experiments to evaluate the effectiveness of the Japanese-English CLIR using the proposed query translation method. The threshold value T_{COT} was set to, max(COT)−4.0, max(COT)×0.7, max(COT)×0.9, and max(COT)×0.8 for COT_{Min}, COT_{LLR}, COT_{DICE}, and COT_{CHI}, respectively (max(COT) is the maximum COT for a query). The threshold value for word occurrence frequency T_{freq} was set to 1/10, 000 of the total number of English documents in the search engine, hence N/10,000. Those values were empirically determined based on the best results in the preliminary experiments.

5.1 Test Data

For the test data, we used NACSIS Test Collection 1 (NTCIR-1)[17] (Research Purpose Use). It contains summaries of papers presented at conferences hosted by 65 Japanese academic societies, and we used E-Collection, which contains about 190,000 English summaries. Although our system is essentially targeting Web documents, we used NTCIR collection because there is no other test collection suitable for evaluating Japanese-English CLIR of Web documents at present. We used 39 Japanese search topics for evaluation and used only TITLE field, which is a very short description of the topic, for queries. It contains 1-7 words (2.7 words on average) and it resembles to the queries often submitted by an end-user of Web search engines in terms of length. TITLE fields are segmented into words using ChaSen[18] morphological analyzer and only nouns and unknown terms were used as a query.

5.2 Language Resources for the Experiments

5.2.1 Bilingual Dictionary

For query translation, we merged three dictionaries, Japanese-English *Bilingual Dictionary* and *Technical Terms Dictionary (Information Processing)* of EDR Electronic Dictionary Version 1.5[19] and EDICT, which is a freeware Japanese-English dictionary. The total vocabulary of the merged dictionaries was 366,041 terms. However, it was not sufficient for translating some of the queries used for the experiments. In order to avoid the effect of the quality of the dictionary, we added translations for 18 words that appeared in the queries.

5.2.2 Monolingual Corpus

We used a Web search engine as a monolingual corpus, as described in Section 4.3. We chose AltaVista as the Web search engine for the query disambiguation for the following reasons:

1. It is possible to specify the target language for the retrieval,
2. It is possible to obtain the total number of documents, which is required to calculate COT,
3. It supports the proximity operator (NEAR) in addition to the AND operator,
4. It has comparatively large index as a Web search engine (about 130 million English documents on March 2000).

Table 3: Results of the experiments using NACSIS Test Collection 1 (Experiment 1).

	NODIS	ONE	AND	NEAR	NTCIR	MAN
0.00	0.4769	0.3987	0.4241	0.4841	0.5051	0.5301
0.20	0.2513	0.2097	0.2427	0.2755	0.2650	0.2778
0.40	0.1732	0.0963	0.1591	0.1677	0.1812	0.1984
0.50	0.1629	0.0871	0.1530	0.1613	0.1745	0.1813
0.70	0.0514	0.0428	0.0536	0.0565	0.0534	0.0526
1.00	0.0031	0.0031	0.0031	0.0033	0.0031	0.0005
Avg.	0.1438	0.1084	0.1371	0.1513	0.1544	0.1564

Generally, Web search engines periodically update their index and the populations of the target collection may fluctuate as time goes on. In the experiments, in order to avoid the effect of such fluctuation, co-occurrence frequencies for a query were obtained continuously within a short period of time and those values were never reused. The number of times querying the search engine for 39 queries was 107.5 on average, but 33.7 on average for queries containing no more than 3 terms (33 queries out of 39). Furthermore, we also experimented with NTCIR-1 E-Collection, which is identical to the target collection, as the monolingual corpus. In this case, the AND operator was used for calculating co-occurrence tendencies.

5.3 Retrieval System

For the retrieval system, we used *Namazu* retrieval system (version 2.0.1). It is a freeware full-text retrieval system based on a Boolean model and supports basic functions such as a composition of Boolean operators, ranking of the results and phrase retrieval.

5.4 Experiment 1 (based on MI)

5.4.1 Experimental Results

Experimental results for *n* term co-occurrence tendency based on MI are shown in Table 3 and Figure 4. In Table, NODIS is the result of no disambiguation, which means using all possible translation-candidates obtained from the dictionary. In this case, all translation-candidates of each term were combined with AND operators, and all translation-candidates were combined with OR operators. ONE is the result of using only one translation-candidate that has the top COT_{MIn}, by using NTCIR collection as a corpus. AND and NEAR are the results of using the proposed disambiguation method described in Section 4.3, which uses the Web search engine as a corpus. If there is no translation-candidate that retrieve documents by using a search engine, all translation-candidates were used. NTCIR

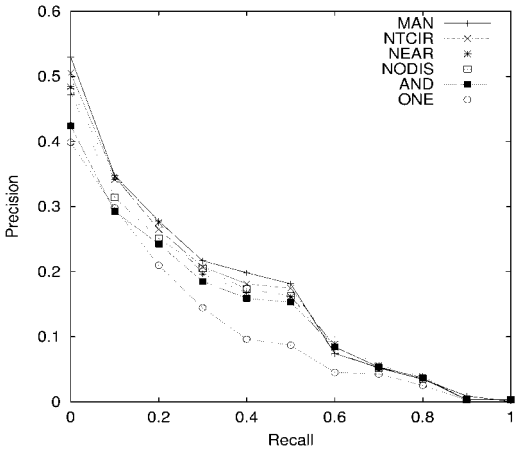

Figure 4: Recall-precision curves of Experiment 1.

is the result of using the proposed disambiguation method, but by using NTCIR collection as a corpus instead of a Web search engine. MAN is the result of English queries manually translated from the original Japanese queries. In the table, the column 0.00-1.00 indicates the precisions at each recall level, and Avg. indicates the average precision.

As an actual example of a disambiguation, the result of NEAR method for the query that consists of two words "神経 (nerve)" and "再生 (regeneration)" (topic ID 0080) is shown in Table 4 and Figure 5. In Table 4, top 7 pairs exceed the threshold value. In this case, the 1st, 3rd, and 4th pairs seem to be the appropriate translations, but all 7 pairs combined by OR operators are used for the actual query.

Table 4: Example of co-occurrence tendency values for a query used in the experiments (top 15 pairs).

Rank	神経	再生	COT_{MI2}
1	**nerve**	**regeneration**	2.20
2	nerve	regrowth	1.82
3	**"nervous system"**	**regeneration**	1.12
4	**nerves**	**regeneration**	0.54
5	nerves	regrowth	0.43
6	"nervous system"	regrowth	-0.17
7	sensation	regrowth	-1.52
8	sensation	reincarnation	-2.33
9	"nervous system"	resuscitation	-2.65
10	sensitivity	playback	-2.95
11	sensitivity	regeneration	-3.07
12	nerve	resuscitation	-3.07
13	sensation	regeneration	-3.09
14	worry	read	-3.27
15	sensation	rebirth	-3.91

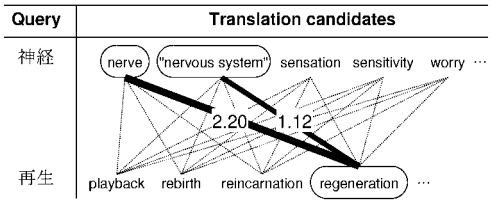

Figure 5: Example of disambiguation for a query used in experiments.

5.4.2 Discussion of the Results

In terms of average precisions compared with NODIS, the proposed disambiguation method improved 1.0 point for NTCIR and 0.8 point for NEAR, but decreased 0.7 point for AND. In general, the effectiveness of using all translation-candidates in a dictionary is about a half of the one using monolingual retrieval[11], but our result of NODIS has greatly surpassed it and has achieved 92% of manual translation (MAN). It is probably because: 1) queries consist mostly of technical terms and their ambiguities were relatively low, 2) query structuring using Boolean operator was much more effective than expected. The result of NOSTR (all translation-candidate terms are combined with OR operators) was 57% of the result of MAN and it is comparable to previous studies.

The result of NTCIR, in which the corpus is identical to the target collection, achieved 99% of MAN and the effect of disambiguation was significant. By using a corpus that is consistent with the target, appropriate translations were selected in most of the cases. However, for the results of using Web documents as a corpus, NEAR achieved 97% of the result of MAN, but AND did 88% of MAN and it was lower than NODIS. It is probably due to the scope of the co-occurrence. In AND, the scope is one document and there are more chances for errors, but in NEAR, the scope is 10 words and relatively a better co-occurrence tendency was acquired. In this experiment, NTCIR achieved the highest performance, but our primary target is the Web CLIR. It is impractical to prepare a comprehensive corpus that covers all possible domains. Therefore, our method using Web documents as a monolingual corpus will be effective for Web CLIR.

On the other hand, the result of using only one translation-candidate that has the highest COT_{MIn} achieved only 69% of MAN. The reason is that the highest-ranked translation-candidate is not always appropriate. In our method, by including all translation-candidates that exceed the threshold value, appropriate translations were selected in most of the cases. For example, in a query which consists of "Zipf" and "法則 (law or rule)", the highest-ranked translation-candidate was the pair ("Zipf", "rule"), but the most appropriate translation ("Zipf", "law") also exceeded the threshold and was selected for the final query. Moreover, the number of appropriate translations is not necessarily one. Another example is the Japanese term "施設", which has translations of similar meaning "facility" and "institution" in English. In some context, both of them might be appropriate. In our method, both of them were selected as an effect of selecting multiple translation-candidates that exceed the threshold.

5.5 Experiment 2 (comparison of COT)

We have conducted a comparative experiment for four co-occurrence tendency measures described in Section 4.1. In this experiment, NTCIR collection was used as a corpus.

5.5.1 Experimental Results

The results of the experiment are shown in Table 5. NODIS and MAN in the figure are the same as experiment 1. DICE, LLR, and CHI are the results of disambiguation using modified Dice coefficient, log likelihood ratio, and χ^2 test, respectively.

Table 5: Results of the comparative experiment about co-occurrence tendency measures (Experiment 2).

	DICE	LLR	CHI
0.00	0.5051	0.4973	0.4956
0.10	0.3464	0.3504	0.3498
0.20	0.2650	0.2651	0.2658
0.30	0.2070	0.2066	0.2073
0.40	0.1820	0.1794	0.1822
0.50	0.1753	0.1722	0.1755
0.60	0.0859	0.0863	0.0862
0.70	0.0546	0.0550	0.0549
0.80	0.0366	0.0370	0.0368
0.90	0.0031	0.0031	0.0031
1.00	0.0031	0.0031	0.0031
Avg. Prec.	0.1546	0.1550	0.1549

5.5.2 Discussion of the Results

In this experiment, differences of the average precisions among four measures were lower than 0.1% and no significant difference was observed. It is probably due to the limitation of the target collection, which is a collection of technical documents in a limited domain. In this case, even if inappropriate translations were included in the final query, they were not likely to appear in the target collection. But for Web CLIR that includes documents in various domains, difference among measures might be significant.

6 Conclusion

In this paper, we proposed a method for query term disambiguation using a Web search engine, which is readily available. The results of the experiments showed that our method is effective for very short queries, which

are often used by an end-user of Web search engines. We also showed that our method can achieve a comparable effectiveness with the manual translation, using a corpus that is consistent with the target collection. Our method can easily be extended to other language pairs by preparing only a dictionary. Besides, disambiguated queries are simple Boolean queries and can be simply fed into an existing Web search engine. Future work include detailed considerations on setting the threshold value and the scope of the co-occurrence, an evaluation using actual Web documents, extension to other language pairs and a consideration on how quality and quantity of the corpus affect the retrieval effectiveness as well as the possibility to apply some smoothing method.

References

[1] Kikui, G. Identifying the coding system and language of on-line documents using statistical language models. *Transactions of IPSJ*, 1997, **38**(12), pp. 2440-2448.

[2] Sugimoto, S., Maeda, A., Dartois, M., Ohta, J., Nakao, S., Sakaguchi, T. and Tabata, K. Experimental studies on an applet-based document viewer for multilingual WWW Documents — Functional Extension of and Lessons Learned from Multilingual HTML. *In Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL'98)*, Lecture Notes in Computer Science 1513, Springer-Verlag, 1998, pp. 199-214.

[3] Jansen, B. J., Spink, A. and Saracevic, T. Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processing & Management*, 2000, **36**(2), pp. 207-227.

[4] Fujii, A. and Ishikawa, T. Cross-language information retrieval for technical documents. *In Proceedings of the Joint ACL SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, pp. 29-37.

[5] Oard, D. W. Alternative approaches for cross-language text retrieval. *In Electronic Working Notes of the AAAI Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[6] Grefenstette, G., editor. Cross-language information retrieval. The Kluwer International Series on Information Retrieval, Vol. 2. Kluwer Academic Publishers, 1998.

[7] Nie, J., Simard, M., Isabelle, P. and Durand, R. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, 1999, pp. 74-81.

[8] Maeda, A. and Uemura, S. Key technologies for multilingual information processing on WWW. *In Proceedings of the Fourth International Symposium on Standardization of Multilingual Information Technology (MLIT-4)*, 1999, pp. 15-25.

[9] Lin, C., Lin, W., Bian, G. and Chen, H. Description of the NTU Japanese-English cross-lingual information retrieval system used for NTCIR workshop. *In Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, 1999, pp. 145-148.

[10] Jang, M., Myaeng, S. H. and Park, S. Y. Using mutual information to resolve query translation ambiguities and query term weighting. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, 1999, pp. 223-229.

[11] Ballesteros, L. and Croft, W. B. Resolving ambiguity for cross-language retrieval. *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, 1998, pp. 64-71.

[12] Fatiha, S., Maeda, A., Yoshikawa, M. and Uemura, S.: Integrating Dictionary-based and Statistical-based Approaches in Cross-Language Information Retrieval, *IPSJ SIG Note*s, 2000-DBS-121/2000-FI-58, 2000, pp. 61–68.

[13] Ikeno, A., Murata, T., Shimohata, S. and Yamamoto, H. Machine translation using the Internet natural language resources. *In Proceedings of World TELECOM99+Interactive99 Forum*, 1999.

[14] Church, K. W. and Hanks, P. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 1990, **16**(1), pp. 22-29.

[15] Kitamura, M. and Matsumoto, Y. Automatic extraction of translation patterns in parallel corpora. *Transactions of IPSJ*, 1997, **38**(4), pp. 727-736. (in Japanese)

[16] Dunning, T. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 1993, **19**(1), pp. 61-74.

[17] Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H., Hidaka, S. and Adachi, J. The NTCIR workshop: the first evaluation workshop on Japanese text retrieval and cross-lingual information retrieval. *In Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages (IRAL'99)*, 1999.

[18] Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H. and Asahara, M. Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition. Technical Report NAIST-IS-TR99013, Nara Institute of Science and Technology, 1999.

[19] Japan Electronic Dictionary Research Institute, Ltd. EDR electronic dictionary version 1.5 technical guide, Technical Report TR2-007, Japan Electronic Dictionary Research Institute, Ltd., 1996.