

# Using co-occurrence tendencies to improve Cross-Language Information Retrieval

Fatiha Sadat

Université du Québec à Montréal  
201 avenue du Président Kennedy, Montréal, Québec, H2X 3Y7, Canada

## Abstract

Query disambiguation is considered as one of the most important methods in improving the effectiveness of information retrieval. In the present paper, we focus on query terms disambiguation via, a combined statistical method both before and after translation, in order to avoid source language ambiguity as well as incorrect selection of target translations. By combining query expansion with dictionary-based translation and statistics-based disambiguation, in order to overcome query terms ambiguity, information retrieval should become much more efficient. Thus, query expansion techniques through relevance feedback were performed prior to either the first or the second disambiguation processes. We tested the effectiveness of the proposed combined method, by an application to a French-English Information Retrieval. Experiments involving TREC data collection revealed the proposed disambiguation and expansion methods to be highly effective.

**Keywords:** Cross-Language Information Retrieval, co-occurrence tendency, disambiguation, expansion

## 1. Introduction

In recent years, the number of studies concerning Cross-Language Information Retrieval (CLIR) has grown rapidly, due to the increased availability of linguistic resources for research. Cross-Language Information Retrieval consists of providing a query in one language and searching document collections in one or more languages. Therefore, a translation form is required. In the present paper, we focus on query translation, disambiguation and expansion in order to improve the effectiveness of information retrieval through various combinations of these methods. First, we are interested to find retrieval methods that are capable of performing across languages and which do not rely on scarce resources such as parallel corpora. Bilingual Machine Readable-Dictionaries (MRDs), more prevalent than parallel texts, appear to be a good alternative. However, simple translations tend to be ambiguous and yield poor results. A combination that includes a statistical approach for a disambiguation can significantly reduce errors associated with *polysemy*<sup>1</sup> in dictionary translation. In addition, automatic query expansion, which has been known to be among the most important methods in overcoming the word mismatch problem in information retrieval, is also considered. As an assumption to reduce the effect of ambiguity and errors that a dictionary-based method would cause, a combined statistical disambiguation method is performed both prior to and after translation. Although, the proposed information retrieval system is general across languages in information retrieval, we conducted experiments and evaluations concerning French-English information retrieval.

---

<sup>1</sup> Polysemy is a word or phrase, which has multiple meanings.

The remainder of the present paper is organized as follows. Section 2 provides a brief overview of related works. Both dictionary-based and the proposed disambiguation methods are described in Section 3. A combination involving query expansion is described in Section 4. Evaluation and discussion of the experiments of the present study are presented in Section 5. Section 6 describes the conclusion of the present paper.

## 2. Related Research in CLIR

The potential of knowledge-based technology has led to increasing interest in CLIR. The query translation of an automatic MRD, on its own, has been found to lead to a drop in effectiveness of 40-60% compared to monolingual retrieval (Hull and Grefenstette, 1996; Ballesteros and Croft, 1998). Previous studies have used MRDs successfully, for query translation and information retrieval (Yamabana et al., 1996; Ballesteros and Croft, 1998; Hull and Grefenstette, 1996). However, two factors limit the performance of this approach. The first is that many words do not have a unique translation and sometimes the alternate translations have very different meanings (*homonymy and polysemy*). The fact that a single word may have more than one sense is called *ambiguity*. Translation ambiguity significantly exacerbates the problem in CLIR (Oard, 1997). Most of the previously proposed disambiguation strategies rely on statistical approaches, but without considering ranking or selection of source query terms, which affect directly the selection of target translations. The second challenge is that dictionary may lack some terms that are essential for a correct interpretation of the query. In the present study, we propose the concept of the combined statistical disambiguation technique, applied prior to and after dictionary translation to solve lexical semantic ambiguity. In addition, a monolingual thesaurus is introduced to overcome bilingual dictionary limitation. Automatic query expansion through relevance feedback, which has been used extensively to improve the effectiveness of an information retrieval (Ballesteros and Croft, 1998; Loupy et al., 1998), is considered. Selection of expansion terms was performed through various means. In the present study, we use a ranking factor to select the best expansion terms-those related to all source query terms, rather than to just one query term.

## 3. Translation/Disambiguation in CLIR

There are two types of lexical semantic ambiguity with which a machine translation system must contend: there is ambiguity in the source language where the meaning of a word is not immediately apparent but also ambiguity in the target language when a word is not ambiguous in the source language but it has two or more possible translations (Hutchins and Sommers, 1992). In the present research, query translation/disambiguation phases are performed after a simple *stemming* process of query terms, replacing each term with its inflectional root and each verb with its infinitive form, as well removing most plural word forms, stop words and stop phrases. Three primary tasks are completed using the translation/disambiguation module. First, an *organization of source query terms*, which is considered key to the success of the disambiguation process, will select best pairs of source query terms. Next a *term-by-term translation* using the dictionary-based method (Sadat et al., 2001), where each term or phrase in the query is replaced by a list of its possible translations, is completed. This may occur either because the query deals with a technical topic, which is outside the scope of the dictionary or because the user has entered some form of abbreviations or slang, which is not included in the dictionary (Oard, 1997). To solve this problem, an automatic *compensation* is introduced, via synonym dictionary or existing thesaurus in the concerned language. This case requires an extra

step to look up the query term in the thesaurus or synonym dictionary, find equivalent terms or synonyms of the targeted source term, thus performing a query translation. In addition, short queries of one term are concerned by this phase. The third task, *disambiguation of target translations*, selects best translations related to each source query term. Finally, documents are retrieved in target language.

Fig. 1 shows the overall design of the proposed information retrieval system. Query expansion will be applied prior to and/or after the translation/disambiguation process. Among the proposed expansion strategies are, relevance feedback and thesaurus-based expansion, which could be interactive or automatic.

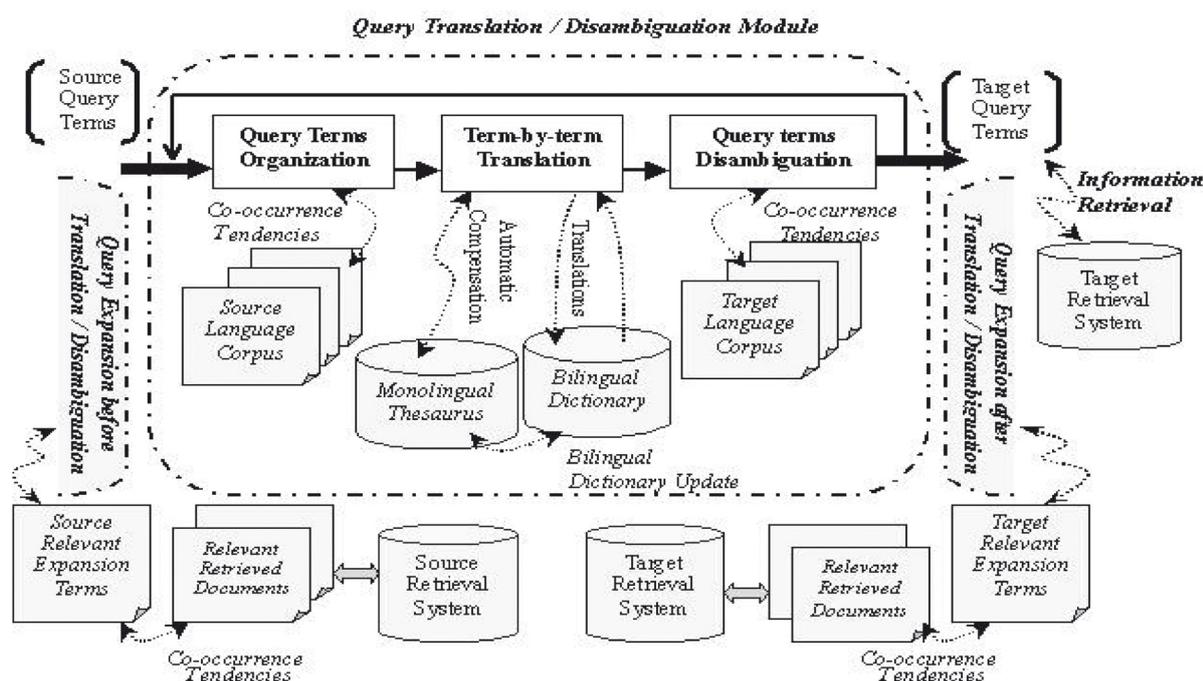


Figure 1: An overview of the Proposed Information Retrieval System (In this research, source/target languages are French/English)

### 3.1. Organization of source query terms

All possible combinations of source query terms are constructed and ranked depending on their mutual co-occurrence in a training corpus. A type of statistical process called *co-occurrence tendency* (Maeda et al., 2000; Sadat et al., 2001) can be used to accomplish this task. Methods such as Mutual Information MI (Church and Hanks, 1990), the Log-Likelihood Ratio LLR (Dunning, 1993), the Modified Dice Coefficient or Gale's method (Gale and Church, 1991) are all candidates to the co-occurrence tendency.

#### 3.1.1. Co-occurrence Tendency

If two elements often co-occur in the corpus, then these elements have a high probability of being the best translations among the candidates for the query terms. The selection of pairs of source query terms to translate as well as the disambiguation of translation candidates in order to select target ones, is performed by applying one of the statistical methods based on co-occurrence tendency, as follows:

- *Mutual Information (MI)*.

This estimation uses *mutual information* as a metric for significance of word co-occurrence tendency (Church and Hanks, 1990), as follows:

$$MI(w_1, w_2) = \log \frac{Prob(w_1, w_2)}{Prob(w_1)Prob(w_2)}$$

Here,  $Prob(w_i)$  is the frequency of occurrence of word  $w_i$  divided by the size of the corpus  $N$ , and  $Prob(w_i, w_j)$  is the frequency of occurrence of both  $w_i$  and  $w_j$  together in a fixed window size in a training corpus, divided by the size of the corpus  $N$ .

- *Log-Likelihood Ratio (LLR)*.

The Log-Likelihood Ratio (Dunning, 1993) has been used in many researches. LLR is expressed as follows:

$$-2\log\lambda = K_{11}\log\frac{K_{11}N}{C_1R_1} + K_{12}\log\frac{K_{12}N}{C_1R_2} + K_{21}\log\frac{K_{21}N}{C_2R_1} + K_{22}\log\frac{K_{22}N}{C_2R_2}$$

Where,

$$C_1 = K_{11} + K_{12}, C_2 = K_{21} + K_{22}, R_1 = K_{11} + K_{21}, R_2 = K_{12} + K_{22}, N = K_{11} + K_{12} + K_{21} + K_{22},$$

$K_{11}$  = frequency of common occurrences of word  $w_i$  and word  $w_j$ ,

$K_{12}$  = corpus frequency of word  $w_i$  -  $K_{11}$ ,

$K_{21}$  = corpus frequency of word  $w_j$  -  $K_{11}$ ,  $K_{22}$  =  $N - K_{12} - K_{21}$ .

### 3.2. Disambiguation of Target Translations

A word is *polysemous* if it has senses that are different but closely related. As a noun, for example, *right* can mean something that is morally acceptable, something that is factually correct, or one's entitlement. A two-terms disambiguation of translation candidates can be applied (Maeda et al., 2000; Sadat et al., 2001) is required, following a dictionary-based method. All source query terms are generated, weighed, ranked and translated for a disambiguation through co-occurrence tendency. The classical procedure for a *two-term disambiguation*, is described as follows:

1. Construct all possible combinations of pairs of terms, from the translation candidates.
2. Request the disambiguation module to obtain the co-occurrence tendencies. The window size is set to one paragraph of a text document rather than a fixed number of words.
3. Choose the translation, which shows the highest co-occurrence tendency, as the most appropriate.

As illustrated in Fig. 2, the disambiguation procedure is used for two-term queries due to the computational cost (Maeda et al., 2000). In addition, the primary problem concerning long queries, involves the selection of pairs of terms, as well as the order for disambiguation. We propose and compare two methods for *n-term disambiguation*, for queries of two or more terms. The first method is based on a ranking of pairs of source query terms before the translation and disambiguation of target translations. The key concept in this step is to maintain the ranking order from the organization phase and perform translation and disambiguation starting from the most informative pair of source terms, i.e. a pair of source query terms having the highest co-occurrence tendency. Co-occurrence tendency is involved in both phases, organization for source language and disambiguation for target language. The second method is based on a ranking of target translation candidates. These methods are described as follows: Suppose,  $Q$  represents a source query with  $n$  terms  $\{s_1, s_2, \dots, s_n\}$ .

First Method: *(Ranking source query terms and disambiguation of target translations)*

1. Construct all possible combinations of terms of one source query:  $(s_1, s_2), (s_1, s_3), \dots (s_{n-1}, s_n)$ .
2. Rank all combinations, according to their co-occurrence tendencies toward highest values.
3. Select the combination  $(s_i, s_j)$ , having the highest co-occurrence tendency, where at least one translation of the source terms has not yet been fixed.
4. Retrieve all related translations to this combination from the bilingual dictionary.
5. Apply a two-term disambiguation process to all possible translation candidates.
6. Fix the best target translations for this combination and discard the other translation candidates.
7. Go to the combination having the next highest co-occurrence tendency, and repeat steps 3 to 7 until every source query term's translation is fixed.

Second Method: *(Ranking and disambiguation of target translations)*

1. Retrieve all possible translation candidates for each source query term  $s_i$ , from the bilingual dictionary.
2. Construct sets of translations  $T_1, T_2, \dots, T_n$  related to each source query term  $s_1, s_2, \dots, s_n$ , and containing all possible translations for the concerned source term. For example,  $T_i = \{t_{i1}, \dots, t_{in}\}$  is the translation set for term  $s_i$ .
3. Construct all possible combinations of elements of different sets of translations. For example,  $(t_{11}, t_{21}), (t_{11}, t_{22}), \dots (t_{ij}, t_{nk})$ .
4. Select the combination having the highest co-occurrence tendency<sup>2</sup>.

Fix these target translations, for the related source terms and discard the other translation candidates.

Go to the next highest co-occurrence tendency and repeat step 4 through 6, until every source query term's translation is fixed.

Examples using the two proposed disambiguation methods are shown in Fig. 3 and Fig. 4, for source English queries and target French translations.

| Query  | Translation Candidates              |
|--------|-------------------------------------|
| Doctor | Ph -Docteur Médecin Doctorat ...    |
| Drug   | Stupéfiant Remède Drogue Médicament |

Figure 2: Two-Term Disambiguation Process

Highest co-occurrence tendencies for combinations of target translation candidates are as follows: *(médecin, médicament), (médecin, remède), (médecin, drogue) ...*

Source French query: "doctor drug". Translated query to English: "médecin médicament".

| Query  | Translation Candidates              |
|--------|-------------------------------------|
| Doctor | Ph -Docteur Médecin Doctorat ...    |
| Drug   | Médicament Drogue Stupéfiant Remède |
| Cure   | Remède Sécher Guérir Fumer ...      |
| Office | Bureau Fonction Cabine ...          |

Figure 3: N-Term Disambiguation (First Method): Ranking Source Query Terms and Disambiguation of Target Translations

Highest co-occurrence tendencies related to pairs of source query terms are as follows: (drug, cure), (doctor, drug), (doctor, office), (doctor, cure)...

Source French query: "doctor drug cure office". Translated query to English: "médecin médicament guérir cabinet".

| Query  | Translation Candidates                |
|--------|---------------------------------------|
| Doctor | Ph -Docteur Médecin Frelater Arranger |
| Drug   | Médicament Drogue Supéfiant Remède .. |
| Cure   | Remède Sécher Guérir Fumer Saler ...  |
| Office | Fonction Bureau Cabinet ...           |

Figure 4: N-Term Disambiguation (Second Method): Ranking and Disambiguation of Target Translations

Highest co-occurrence tendencies related to target translation candidates are as follows: (médecin, guérir), (guérir, remède), (remède, médecin) (médecin, fonction)...

Source French query: "doctor drug cure office". Translated query to English: "médecin remède guérir fonction".

#### 4. Query Expansion in CLIR

Following the research reported by (Ballesteros and Croft, 1998) on the use of local feedback, the addition of terms that emphasize query concepts in the pre and post-translation phases improves both precision and recall. In the present study, we have proposed the combined automatic query expansion before and after translation through a relevance feedback. Original queries were modified, using judgments of the relevance of a few highly ranked documents, obtained by an initial retrieval, based on the presumption that those documents are relevant.

However, query expansion must be handled very carefully. Simply selecting any expansion term from relevant retrieved documents could be risky. Therefore, our selection is based on the co-occurrence tendency in conjunction with all terms in the original query, rather than with just one query term. Assume that we have a query  $Q$  with  $n$  terms,  $\{term_1 \dots term_n\}$ , then a ranking factor based on the co-occurrence frequency between each term in the query and an expansion term candidate, already extracted from the top retrieved relevant documents, is evaluated as:

$$Rank(expterm) = \sum_{i=1}^n co-occur(term_i, expterm)$$

where,  $co-occur(term_i, expterm)$  represents the co-occurrence tendency between a query term  $term_i$  and the targeted expansion candidate  $expterm$ .  $Co-occur(term_i, expterm)$  can be evaluated by any estimation technique, such as mutual information or the log-likelihood ratio. All co-occurrence values are computed and then summed for all query terms ( $i=1 \dots n$ ). An expansion candidate having the highest rank is selected as an expansion term for the query  $Q$ . Note that the highest rank must be related to at least the maximum number of terms in the query, if not all query terms. Such expansion may involve several expansion candidates or just a subset of the expansion candidates.

## 5. Experiments and Evaluation

Experiments to evaluate the effectiveness of the two proposed disambiguation strategies, as well as the query expansion, were performed using an application of French-English information retrieval, i.e. French queries to retrieve English documents.

### 5.1. Linguistics Resources

*Test Data:* In the present study, we used test collection 1 from the TREC <sup>2</sup> data collection. Topics 63-150 were considered as English queries and were composed of several fields. Tags <num>, <dom>, <title>, <desc>, <smry>, <narr> and <con> denote topic number, domain, title, description, summary, narrative and concepts fields, respectively. Key terms contained in the title field <title> and description field <desc>, an average of 5.7 terms per query, were used to generate English queries. Original French queries were constructed by a native speaker, using manual translation.

*Monolingual Corpora:* The Canadian Hansard corpus (parliament debates) is a bilingual French-English parallel corpus, which contains more than 100 million words of English text as well as the corresponding French translations. In the present study, we used Hansard as a monolingual corpus for both French and English languages.

*Bilingual Dictionary:* COLLINS French-English dictionary was used for the translation of source queries.

*Monolingual Thesaurus:* EuroWordNet (Vossen, 1998) a lexical database was used to compensate for possible limitations in the bilingual dictionary.

*Stemmer and Stop Words:* Stemming was performed using the English *Porter* <sup>3</sup> *Stemmer*. A special French stemming was developed and used in these experiments.

*Retrieval System:* The *SMART Information Retrieval System* <sup>4</sup> was used to retrieve both English and French documents. SMART is a vector model, which has been used in several studies concerning Cross-Language Information Retrieval.

### 5.2. Experiments and Results

A retrieval using original French/English queries was represented by *Mono\_Fr/Mono\_Eng* methods, respectively. We conducted two types of experiments. Those related to the query translation/disambiguation and those related to the query expansion before and/or after translation. Document retrieval was performed using original and constructed queries by the following methods. *All\_Tr* is the result of using all possible translations for each source query term, obtained from the bilingual dictionary. *No\_DIS* is the result of no disambiguation, which means selecting the first translation as the target translation for each source query term. We tested and evaluated two methods fulfilling the disambiguation of translated queries (after translation) and the organization of source queries (before translation), using the co-occurrence tendency and the following estimations: Log-Likelihood Ratio (LLR) and Mutual Information (MI). LLR was used for *Bi\_DIS*, disambiguation of consecutive pairs of source terms, without any ranking or selection (Sadat et al., 2001), for *LLR\_DIS.bef*, the result of the first proposed disambiguation

---

<sup>2</sup> <http://trec.nist.gov/data.html>.

<sup>3</sup> <http://bogart.sip.ucm.es/cgi-bin/webstem/stem>.

<sup>4</sup> <ftp://ftp.cs.cornell.edu/pub/smart>.

method (ranking source query terms, translation and disambiguation of target translations) and *LLR\_DIS.aft*, the result of the second proposed disambiguation method (ranking and selecting target translation). In addition, MI estimation was applied to *MI\_DIS.bef* and *MI\_DIS.aft*, for the first and second proposed disambiguation methods. Query expansion was completed by the following methods: *Feed.bef\_LLR*, which represents the result of adding a number of terms to the original queries and then performing a translation and disambiguation via *LLR\_DIS.bef*. *Feed.aft*, is the result of query translation, disambiguation via *LLR\_DIS.bef* method and then expansion. Finally, *Feed.bef\_aft*, is the result of combined query expansion both before and after the translation and disambiguation via *LLR\_DIS.bef*. In addition, we tested a query expansion before and after the disambiguation method *MI\_DIS.bef*, together with the following methods: *Feed.bef\_MI*, *Feed.aft\_MI* and *Feed.bef\_aft\_MI*. Results and performance of these methods are described in Tab. 1. Fig. 5 and Fig. 6 show the query translation/disambiguation using LLR and MI. Fig. 7 and Fig. 8 show the query expansion for different combinations and estimations for the co-occurrence tendency (LLR or MI).

### 5.3. Discussion

The first column of Tab. 1 indicates the method. The second column indicates the number of retrieved relevant documents, and the third column indicates the precision averaged at point 0.10 on the Recall/Precision curve. The fourth column is the average precision, which is used as a basis for the evaluation. The fifth column is the R-precision and the sixth column represents the difference in term of average precision of the monolingual counterpart.

|                   | <i>Rel Docs</i> | <i>at 0.10</i> | <i>A. Prec</i> | <i>R. Prec</i> | <i>% Mono</i> |
|-------------------|-----------------|----------------|----------------|----------------|---------------|
| Mono_Fr (origin)  | 434             | 2,9014         | 1,8257         | 2,0313         | 100,00        |
| Mono_Eng (origin) | <b>433</b>      | <b>3,0813</b>  | <b>0,1819</b>  | <b>1,8493</b>  | <b>100,00</b> |
| All_Tr            | 406             | 2,9757         | 1,5000         | 1,7868         | 3,43          |
| No_DIS            | 429             | 2,8674         | 1,5375         | 1,6882         | 3,52          |
| Bi_DIS            | 418             | 2,8576         | 1,5688         | 1,9229         | 85,95         |
| LLR_DIS.Aft       | 431             | 3,4806         | 1,6576         | 1,9535         | 90,82         |
| LLR_DIS.Bef       | <b>434</b>      | <b>3,5722</b>  | <b>1,8604</b>  | <b>2,1653</b>  | <b>101,94</b> |
| MI_DIS.Aft        | 414             | 3,1299         | 1,6146         | 1,7750         | 3,70          |
| MI_DIS.Bef        | 429             | 3,5590         | 1,8417         | 2,1639         | 100,91        |
| Feed.bef_LLR      | 413             | 3,1299         | 1,6035         | 1,8007         | 87,86         |
| Feed.aft_LLR      | 433             | 3,5785         | 1,8493         | 2,1979         | 4,23          |
| Feed.bef_aft_LLR  | <b>436</b>      | <b>3,6403</b>  | <b>1,8778</b>  | <b>2,2229</b>  | <b>102,89</b> |
| Feed.bef_MI       | 405             | 3,0514         | 1,5722         | 1,7507         | 3,59          |
| Feed.aft_MI       | 430             | 3,5646         | 1,8479         | 2,1347         | 4,23          |
| Feed.bef_aft_MI   | <b>430</b>      | <b>3,5833</b>  | <b>1,8896</b>  | <b>2,1368</b>  | <b>4,33</b>   |

Table 1: Evaluations of the Translation, Disambiguation and Expansion Methods (Different combinations with LLR and MI co-occurrence frequencies)

Compared to the retrieval using original queries (English or French), *All\_Tr* and *No\_DIS* showed no improvement in term of precision, recall or average precision, whereas the simple two-term disambiguation *Bi\_DIS* (disambiguation of consecutive pairs of source query terms) has increased the recall, precision and average precision by +1.71% compared to the simple dictionary translation without any disambiguation. On the other hand, the first proposed disambiguation method (ranking and selecting target translations) *LLR\_DIS.aft*, showed a potential precision enhancement, 0.5012 at 0.10 and 90.82% average precision; however, recall was not improved (4.131 relevant documents retrieved). The best performance for the disambiguation process

was achieved by the second proposed disambiguation method (ranking source query terms and selecting target translations) *LLR\_DIS.bef*, in average precision, precision and recall. The average precision was 101.51% of the monolingual counterpart, precision was 0.5144 at 0.10 and 436 relevant documents were retrieved.

This suggests that ranking and selecting pairs for source query terms, is very helpful in the disambiguation process to select best target translations, especially for long queries of at least three terms. Results based on mutual information were less efficient compared to those using log-likelihood ratio. However, ranking source query terms before the translation and disambiguation resulted in an improvement in average precision, 100.91% of the monolingual counterpart. Although, query expansion before translation via *Feed.bef\_LLR/Feed.bef\_MI*, gave an improvement in average precision compared to the non-disambiguation method *No\_DIS*, a slight drop in precision (0.4507/0.4394) and recall (413/405 relevant retrieved documents) was observed compared to *LLR\_DIS.bef* or *MI\_DIS.bef*. However, *Feed.aft\_LLR/Feed.aft\_MI* showed an improvement in average precision, 101.33%/101.25% compared to the monolingual counterpart and improved the precision (0.5153/0.5133 at 0.10) and the recall (433/ 430 retrieved relevant documents). Combined feedbacks both before and after translation yielded the best result, with an improvement in precision (0.5242 at 0.10), recall (434 retrieved relevant documents) and average precision, 102.89% of the monolingual counterpart when using LLR estimation. A disambiguation using MI for co-occurrence tendency yielded a good result, 103.53% of the monolingual counterpart for average precision. These results suggest that combined query expansion both before and after the proposed translation/disambiguation method improves the effectiveness of an information retrieval, when using a co-occurrence tendency based on MI or LLR.

Thus, techniques of primary importance to this successful method can be summarized as follows:

- A statistical disambiguation method based on the co-occurrence tendency was applied first prior to translation, in order to eliminate misleading pairs of terms for translation and disambiguation. Then after translation, the statistical disambiguation method was applied in order to avoid incorrect sense disambiguation and to select best target translations.
- Ranking and careful selection are fundamental to the success of the query translation, when using statistical disambiguation methods.
- A combined statistical disambiguation method before and after translation provides a valuable resource for query translation and thus information retrieval.
- Log-Likelihood Ratio was found to be more efficient for query disambiguation than Mutual Information.
- A co-occurrence measure to select an expansion term was evaluated using all terms of the original query, rather than using just one query term.
- Each type of query expansion has different characteristics and therefore combining various types of query expansion could provide a valuable resource for use in query expansion. This technique offered the greatest performance in average precision.
- These results showed that CLIR could outperform the monolingual retrieval. The intuition of combining different methods for query disambiguation and expansion, before and after translation, has confirmed that monolingual performance is not necessarily the upper bound for CLIR performance (Gao et al., 2001). One reason is that those methods have completed each other and that the proposed query disambiguation had a positive effect during the translation and thus retrieval. Combination to query expansion had an effect on the translation as well, because related words could be added.

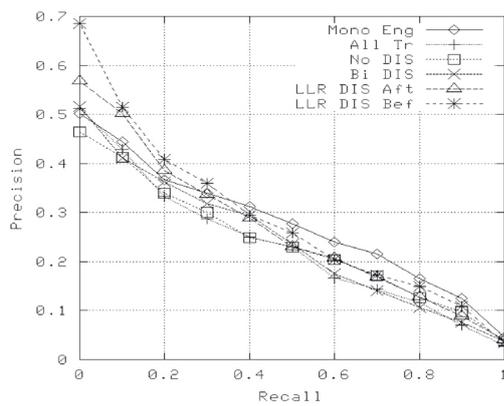


Figure 5: Recall/Precision Curves for the Query Translation/Disambiguation using LLR estimation

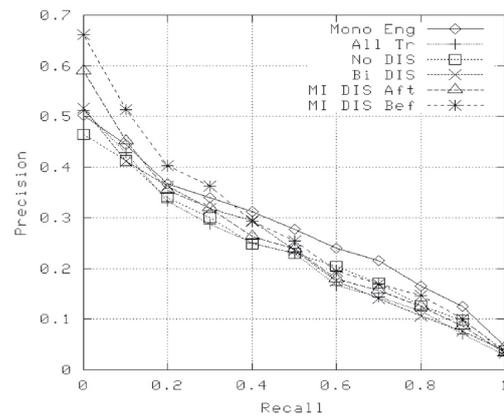


Figure 6: Recall/Precision Curves for the Query Translation/Disambiguation using MI estimation

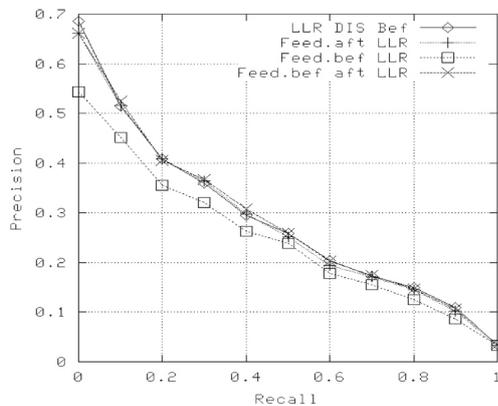


Figure 7: Recall/Precision Curves for the Query Expansion before and after the Translation/ Disambiguation using LLR estimation

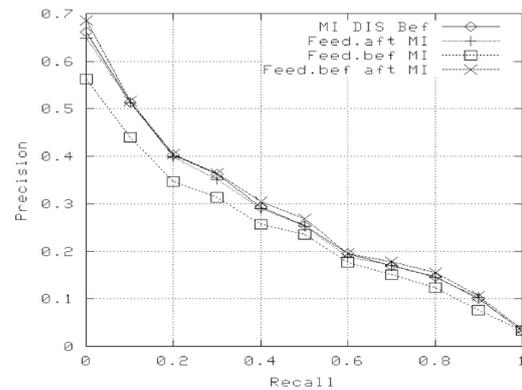


Figure 8: Recall/Precision Curves for the Query Expansion before and after the Translation/ Disambiguation using MI estimation

The proposed combined disambiguation method prior to and after translation, was based on a selection of one target translation in order to retrieve documents. Setting a threshold in order to select more than one target translation is possible using weighting scheme for the selected target translations in order to eliminate misleading terms and construct an optimal query to retrieve documents.

## 6. Conclusion

Dictionary-based method is attractive for several reasons. This method is cost effective and easy to perform, resources are readily available and performance is similar to that of other Cross-Language Information Retrieval methods. Ambiguity arising from failure to translate queries is largely responsible for large drops in effectiveness below monolingual performance (Ballesteros and Croft, 1998). The proposed disambiguation approach of using statistical information from language corpora to overcome limitation of simple word-by-word dictionary-based translation has proved its effectiveness, in the context of information retrieval. A co-occurrence tendency based on a log-likelihood ratio has showed to be more efficient than the one based on mutual

information. The combination of query expansion techniques, both before and after translation through relevance feedback improves the effectiveness of simple word-by-word dictionary translation. We believe that the proposed disambiguation and expansion methods will be useful for simple and efficient retrieval of information across languages.

Ongoing research includes a search for additional methods that may be used to improve the effectiveness of information retrieval. Such methods may include the combination of different resources and techniques for optimal query expansion across languages. In addition, thesauri and relevance feedbacks will be studied in greater depth. A good word sense disambiguation model will incorporate several types of data source that complete each other, such as a part-of-speech tagger into statistical models. Finally, an approach to learning from documents categorization and classification in order to extract relevant expansion terms will be examined in the future.

## References

- Ballesteros L. and Croft W.B. (1998). Resolving Ambiguity for Cross-Language Retrieval. In *Proceedings of the 21<sup>st</sup> ACM SIGIR Conference*, pp. 64-71.
- Church K.W. and Hanks P. (1990). Word association Norms, Mutual Information and Lexicography. *Computational Linguistics*, vol. 16 (1): 22-29.
- Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, vol. 19 (1): 61-74.
- Gale W.A. and Church K. (1991). Identifying word correspondences in parallel texts. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pp. 152-157.
- Gao J., Nie J.Y., Xun E., Zhang J. Zhou and Huang M.C. (2001). Improving query translation for Cross-Language Information Retrieval using statistical models. In *Proceedings of the 24<sup>th</sup> ACM SIGIR Conference*, pp. 96-104.
- Hull D. (1998). A weighted boolean model for Cross-Language text Retrieval. In Grefenstette, G., editor, *Cross-Language Information Retrieval, chapter 10*. Dordrecht: Kluwer Academic Publishers.
- Hull D. and Grefenstette G. (1996). Querying across languages. A dictionary-based approach to Multilingual Information Retrieval. In *Proceedings of the 19<sup>th</sup> ACM SIGIR Conference*, pp. 49-57.
- Hutchins J. and Sommers J. (1992). *Introduction to Machine Translation*. London: Academic Press.
- Krovetz R. and Croft W. (1992). Lexical ambiguity and information retrieval. In *ACM Transactions on Information Systems*, 10 (2): 115-141.
- Loupy C., Bellot P., El-Beze M. and Marteau P.-F. (1998). Query expansion and classification of retrieved documents. In *Proceedings of TREC-7. NIST Special Publication*.
- Maeda A., Sadat F., Yoshikawa M. and Uemura S. (2000). Query term disambiguation for Web Cross-Language Information Retrieval using a search engine. In *Proceedings of the 5<sup>th</sup> International Workshop on Information Retrieval with Asian Languages*, pp. 25-32.
- Oard D.W. (1997). Alternative approaches for Cross-Language Information Retrieval. In *Working notes of the AAAI Symposium on Cross-Language Text and Speech Retrieval*, Stanford University, USA. <http://www.glue.umd.edu/~oard/research.html>.
- Sadat F., Maeda A., Yoshikawa M. and Uemura S. (2001). Query expansion techniques for the CLEF bilingual track. In *Working Notes for the CLEF 2001 Workshop*, pp. 99-104.

Vossen P. (1998). *EuroWordNet. A Multilingual database with lexical semantic networks*. Dordrecht: Kluwer Academic Publishers.

Yamabana K., Muraki K., Doi S. and Kamei S. (1996). A language conversion Front-End for Cross-Linguistic Information Retrieval. In *Proceedings of SIGIR Workshop on CLIR*, Zurich, Switzerland, pp. 34-39.