Resolving Ambiguity for Cross-language Retrieval

Lisa Ballesteros
balleste@cs.umass.edu
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts
Amherst, MA 01003-4610 USA
http://ciir.cs.umass.edu/

W. Bruce Croft
croftcs.umass.edu
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts
Amherst, MA 01003-4610 USA
http://ciir.cs.umass.edu/

Abstract One of the main hurdles to improved CLIR effectiveness is resolving ambiguity associated with translation. Availability of resources is also a problem. First we present a technique based on co-occurrence statistics from unlinked corpora which can be used to reduce the ambiguity associated with phrasal and term translation. We then combine this method with other techniques for reducing ambiguity and achieve more than 90% monolingual effectiveness. Finally, we compare the co-occurrence method with parallel corpus and machine translation techniques and show that good retrieval effectiveness can be achieved without complex resources.

1 Introduction

Research in the area of cross-language information retrieval (CLIR) has focused mainly on methods for translating queries. Full document translation for large collections is impractical, thus query translation is a viable alternative. Methods for translation have focused on three areas: *dictionary translation*, *parallel* or *comparable corpora* for generating a translation model, and the employment of *machine translation* (MT) techniques. Despite promising experimental results with each of these approaches, the main hurdle to improved CLIR effectiveness is resolving ambiguity associated with translation.

In addition to the ambiguity problem, each of the approaches to CLIR has drawbacks associated with the availability of resources. This is made more critical as the number of languages represented in electronic media continues to expand. MT systems can be employed [GLY96], but tend to need more context than is in a query for accurate translation. The development of such a system requires an enormous amount of time and resources. Even if a system works well for one pair of languages, each new language pair requires a significant new effort. Parallel corpora are being used by several groups e.g.[LL90, Dav96, CYF$^+$97]. One approach at NMSU [DO97] has been to translate via machine readable dictionaries (MRD) followed by a disambiguation phase using part-of-speech (POS) and parallel corpus analysis. However, parallel corpora are hard to come by. They tend also to have narrow coverage and may not yield the level of disambiguation necessary in a more general domain. Work at ETH has focused [SB96] on using comparable corpora to build similarity thesauri which generate a translation effect. This method has been shown to be especially effective when the corpora are domain specific [SBS97]. Comparable corpora although not direct translations, contain documents matched by topic. However, it is not clear that they are easier to construct than are parallel document collections. As with parallel corpora, the question remains of what other disambiguation methods could be used in a more general context to augment these techniques. Dictionary translation has been the starting point for other researchers [BC96, HG96]. The method relies on the availability of machine readable dictionaries (MRD). Dictionaries like the other resources mentioned, may be proprietary or costly. Although on-line dictionaries are becoming more widely available, the coverage and quality may be lower than one would like.

Regardless of the cross-language approach taken, translation ambiguity is a problem which must be addressed. Resources for cross-language retrieval can require tremendous manual effort to generate and may be difficult to acquire. Therefore methods which capitalize on existing resources must be found. In this paper, we describe a technique that employs co-occurrence statistics obtained from the corpus being searched to disambiguate dictionary translations. We focus on the *translation of phrases* which has been shown to be especially problematic. We also explore the disambiguation of term translations. Finally, we compare the effectiveness of the *co-occurrence method* with that of several others: *parallel corpus disambiguation*; word and phrase *dictionary translation augmented by query expansion* at various stages of the translation process; and two *machine translation* systems. Results show that co-occurrence statistics can successfully be used to reduce translation ambiguity.

2 Dictionary Translation and Ambiguity

Cross-language effectiveness using MRD's can be more than 60% below that of mono-lingual retrieval. Simple dictionary translation via machine readable dictionary yields ambiguous translations. Target language queries are translated by replacing source language words or multi-term concepts by their target language equivalents. Translation error is due to three factors [BC96, HG96]. The first factor is the addition of extraneous terms to the query. This is because a dictionary entry may list several senses for a term, each having one or more possible translations. The second is failure to translate technical terminology which is often not found in general dictionaries. Third is the failure to translate multi-term concepts as phrases or to translate them poorly. Previous work [BC97] showed how query expansion could be used to reduce translation error and bring cross-language effectiveness up to 68% of monolingual. However, this still leaves a lot of room for improvement.

Our hypothesis is that the correct translations of query terms will co-occur as part of a sub-language and that incorrect translations will tend not to co-occur. This information could be used to translate compositional phrases, thus reducing the ambiguity associated with word-by-word translation. Additionally, we propose that disambiguation methods using unlinked corpora can be as effective as those using parallel or comparable corpora. The details of the parallel corpus method and the proposed co-occurrence method are given in the next sections.

2.1 Parallel Corpus Disambiguation

Parallel corpora contain a set of documents and their translations in one or more other languages. Analysis of these paired documents can be used to infer the most likely translations of terms between languages in the corpus. We employ parallel corpus analysis to look at the impact of query term disambiguation on CLIR effectiveness. The technique is a modification of one used by NMSU [DO97] and is described below.

Source language (Spanish) queries are first tagged using a part-of-speech (POS) tagger. Each Spanish source term is replaced by all possible target language (English) translations for the term's POS. If there is no translation corresponding to a particular query term's tag, the translations for all parts-of-speech listed in the dictionary for that term are returned. There may be one or more ways to translate a given term. When more than one equivalent is returned, the best single term is chosen via parallel corpus disambiguation.

Disambiguation proceeds in the following way. The top 30 Spanish documents are retrieved from the parallel UN corpus in response to a Spanish query. The top 5000 terms based on Rocchio ranking are extracted from the English UN documents that correspond to the top 30 Spanish documents. The translations of a query term are ranked by their score in the list of 5000. The highest ranking translation(s) is chosen as the "best“ translation for that term. If none of the equivalents are on the list, no disambiguation is performed and all equivalents are chosen. This method differs from that of NMSU in two ways. First, we used document level alignment instead of sentence level alignment. Second, rather than disambiguation based on the top documents retrieved in response to the query, they retrieved the top sentences in response to a query term. They then chose the term translation that retrieved the most sentences like those retrieved for the untranslated term.

2.2 Disambiguation using Co-occurrence Statistics

The correct translations of query terms should co-occur in target language documents and incorrect translations should tend not to co-occur. We use this hypothesis as the foundation for a method to disambiguate phrase translations. Given the possible target equivalents for two source terms, we infer the most likely translations by looking at the pattern of co-occurrence for each possible pair of definitions. Co-occurrence statistics have been used with some success for phrasal translations [SMH96, Kup93]. These techniques rely on parallel corpora and our interest is in ascertaining whether unlinked corpora can be used effectively for phrasal translation. Kraaij and Hiemstra [KH97] used co-occurrence frequency for phrase translation with some success during the TREC-6 [Har97] evaluations. In [DIS91] a co-occurrence method was used for target word selection, however there have been no reports of its use in a retrieval environment. A description of our method follows.

Given two tagged source terms, collect all target translation equivalents appropriate to each term's part-of-speech. Generate all possible sets $\{a, b\}$ such that a is a definition of $term1$ and b is a definition of $term2$. Measure the importance of co-occurrence of the elements in a set by the *em* metric [XC98]. It is a variation of EMIM [vR77] and measures the percentage of the occurrences of a and b which are net co-occurrences (co-occurrences minus expected co-occurrences), but unlike EMIM does not favor uncommon co-occurrences.

$$em(a,b) = max(\frac{n_{ab} - En(a,b)}{n_a + n_b}, 0)$$

where n_a, n_b are the number of occurrences of a and b in the corpus, and n_{ab} is the number of times both a and b fall in a text window of t words. $En(a,b) = \frac{n_a n_b}{N}$ and N is the number of text windows in the corpus. Each set is ranked by em score and the highest ranking set is taken as the appropriate translation. If more than one set has a rank of one, all of them are taken as translations. Our method differs from that of Dagan, et al. in the following ways. They paired words to be translated via syntactic relationships e.g. subject-verb. Selection was made via a statistical model based on the ratio of the frequency of co-occurrence for one alternative versus the frequency of co-occurrence of all the alternatives.

3 Experiments

Word-by-word dictionary translations are error prone for the reasons given in section 2. In this paper, we explore several methods for disambiguating dictionary-based query translations. We focus on phrase translations and demonstrate the effectiveness of a disambiguation method based on co-occurrence statistics (CO) gathered from unlinked corpora. We also show that term translations may be disambiguated via co-occurrence analysis. CO is compared to a disambiguation technique based on parallel corpora (PLC). These methods are combined with other techniques for reducing ambiguity and a comparison of their effectiveness with that of query translation via machine translation is given. Our experiments are described in more detail below.

The experiments in this study were limited to one language pair. Spanish (source language) queries were translated to English (target language). The queries consisted of twenty-one TREC Cross-language topics with an average of 7.6 non-stopwords per query. Table 1 gives sample queries and their correct translations. Evaluation was performed on the 748 MB TREC AP English collection (having 243K documents covering '88-'90) with provided relevance judgments. Co-occurrence statistics were collected from the portion of the AP collection covering 1989. This dataset is a first-time collection with pooled relevance judgments from thirteen retrieval systems. However, the preliminary nature of the data shouldn't greatly effect the outcome of our experiments.

Queries were processed in the following way. First, queries were tagged by a part-of-speech (POS) tagger. Sequences of nouns and adjective-noun pairs were taken to be phrases. Automatic translations were performed by translating phrases as multi-term concepts when possible and individual terms word-by-word. Stop words and stop phrases such as "A relevant document will" were also removed.

The word-by-word translations were performed by replacing query terms in the source language with the dictionary definition of those terms in the target language. Term translations were disambiguated by transferring only those definitions matching a query term's POS. When more than one translation existed for a term, they were all wrapped in an INQUERY #synonym operator. Words that were not found in the dictionary were added to the new query without translation. The Collins Spanish-English bilingual MRD was used for the translations. For a more detailed description of this process, see [BC96]. Section 4 compares the effectiveness of disambiguating term

Caso Waldeheim. Razones de la controversia que rodea las acciones de Waldheim durante la Segunda Guerra Mundial. Waldheim Case. Reasons for the controversy surrounding the actions of Waldheim during the Second World War.
Educación sexual. El uso de la educación sexual para combatir el SIDA. Sex Education. The use of sex education to combat AIDS.
Fast food in Europe. How successful is the spread of American fast food franchises in Europe? Comida rápida en Europa. Qué tan exitosa ha sido la expansión de concesiones americanas en Europa?

Table 1: Three Spanish queries with English translations.

translations via POS and the #synonym operator with word-by-word translation without disambiguation.

Phrasal translations were performed using information on phrases and word usage contained in the Collins MRD. This allowed the replacement of a source phrase with its multi-term representation in the target language. When a phrase could not be defined using this information, the remaining phrase terms were translated in one of two ways. Terms were translated word-by-word followed by parallel corpus disambiguation (PLC) described in section 2.1, or they were translated as multi-term concepts using the co-occurrence method (CO) described in section 2.2. Recall that PLC disambiguates terms using the entire query as context, while the CO method only uses the context of a phrasal unit. All CO experiments were run with a text window size of 250 terms. Section 5, compares the ability of the CO method with that of the phrase dictionary alone for translating phrases. The types of phrases translated and the effectiveness of the methods are given. Section 6 compares disambiguation of term translations via CO with disambiguation via PLC. We also compare the effectiveness of CO and PLC for reducing the error caused by failure to translate phrases as multi-term concepts.

Query expansion before or after automatic translation via MRD significantly reduces translation error. Pre-translation expansion creates a stronger base for translation and improves precision. Expansion after MRD translation introduces terms which de-emphasize irrelevant translations to reduce ambiguity and improve recall. Combining pre- and post-translation expansion increases both precision and recall. Improvement appears to be due to the removal of error caused by the addition of extraneous terms via the translation process.

Section 7 reports on the effectiveness of combining disambiguation methods described above with query expansion which was shown to reduce translation ambiguity in [BC97, BC96]. Query expansion was done via Local Context Analysis (LCA) which is described more fully in [XC96]. LCA is a modification of local feedback [AF77]. It differs from local feedback in that the query is expanded with the best concepts from the top ranked passages rather than the top ranked documents. Training data for the pre-translation LCA experiments consisted of the documents in the 208 MB El Norte (ISM) database from the TREC collection.

Non-interpolated average precision on the top 1000 retrieved documents is used as the basis of evaluation for all experiments. We also report precision at five, ten, twenty, thirty, and one-hundred documents retrieved. All work in this study was performed using the INQUERY information retrieval system. INQUERY is based on the Bayesian inference net model and is described in [TC91b, TC91a, CCB95]. All significance tests used the paired sign test.

4 Disambiguating Word-By-Word Translations

If each source language term has more than one target language equivalent, its term translations will be ambiguous. In these experiments, queries were translated word-by-word and we demonstrate the disambiguating effect of two simple techniques. First, we reduce the number of target language equivalents by replacing each source term with only those equivalents corresponding to a term's part-of-speech. Second, we wrap a #synonym operator around term translations having more than one target term equivalent. If the synonym operator is not used, infrequent terms tend to get higher belief values due to their high idf. The operator treats occurrences of all words within it as occurrences of a single pseudo-term whose document frequency (df) is the sum of df's for each word in the operator. This de-emphasizes infrequent words and has a disambiguation effect.

Table 2 shows the positive effect on average precision for both techniques. Column one corresponds to a word-by-word translation (WBW) of all queries with no attempt at disambiguation. Column two shows the effect of the synonym operator on WBW. Column three shows a word-by-word translation using only POS to disambiguate. The last column combines the disambiguation effects of POS tagging and the use of the synonym operator.

Query	WBW	SYN	POS	POS+SYN
Avg.Prec.	0.1234	0.1784	0.1504	0.2331
% change		44.6	21.9	89.0
Precision at:				
5 docs:	0.2286	0.2762	0.3048	0.3619
10 docs:	0.2286	0.2381	0.3000	0.3286
20 docs:	0.1929	0.2190	0.2476	0.3095
30 docs:	0.1667	0.1968	0.2286	0.2810
100 docs:	0.0786	0.1129	0.1362	0.1705

Table 2: Average precision for word-by-word translation, word-by-word translation augmented by POS disambiguation, synonym operator disambiguation, and word-by-word translation augmented by POS and synonym operator disambiguation.

The synonym operator is more effective for disambiguating than is part-of-speech, with the former primarily affecting precision and the later primarily affecting recall. Combining the two techniques is most effective and greatly improves both precision and recall.

5 Disambiguating Phrasal Translations

As mentioned above, translating multi-term concepts as phrases is an important step in reducing translation error. In these experiments, we compare the ability of our phrase dictionary with that of the co-occurrence method (CO) (as described in 2.2) to translate phrases. We then use co-occurrence statistics to reduce ambiguity by inferring the correct translation of phrases not translatable via our phrase dictionary and compare the effectiveness of the two methods with word-by-word translation as a baseline.

Given the phrases in our query set, we compared the number for which translations could be found in the phrase dictionary with those translatable via CO. The comparison was done by a human assessor who determined whether phrasal translations via either method were correct. Thirty-three phrases were identified in seventeen out of twenty-one TREC6 queries. Ten phrases were duplicates leaving only twenty-three unique phrases. Table 3 gives statistics for the types of phrases identified and also gives results of the comparison. The first row shows the number and types of phrases. The second and third rows show the numbers of phrases of each type that are translatable via our phrase dictionary and co-occurrence method respectively.

	Unique	Compositional	Non-compositional
	23	21	2
Phr. Dict	8	6	2
Co-occur.	13	13	N/A

Table 3: Breakdown of total number of phrases and phrase types in queries, including the numbers translatable via phrase dictionary or co-occurrence method.

Translations of phrases found in the phrase dictionary are good. Note that the six compositional phrases found in the phrase dictionary can also be correctly translated via CO. CO will only work for the translation of compositional phrases. For example, the Spanish phrase *medio oriente* is compositional as it can be translated word-by-word as *middle east*. However,the phrase *contaminación del aire* can not be translated compositionally to *air pollution* since *pollution* is not a translation of *contaminación*. Therefore, we rely upon our phrase dictionary for the translation of non-compositional phrases.

Thirteen compositional phrases are translated correctly using the co-occurrence method. For example, *abuso infantil, comercio marfil, proceso paz* are correctly translated to *child abuse, ivory trade, and peace process*, respectively. The possible translation sets for *processo paz* can be generated from the translations of the constituent terms. The target equivalents of *proceso* and *paz* are *process, lapse of time, trial, prosecution, action, lawsuit, proceedings, processing* and *peace, peacefulness, tranquility, peace, peace treaty, kiss of peace, sign of peace*, respectively. The translation of one of the thirteen is not ambiguous since both constituent source terms have only one target translation.

Seven other compositional phrases were not in the phrase dictionary and were translated incorrectly via CO. In these cases, the translation failure does not appear to be a big problem since only one of the queries containing a poorly translated phrase loses effectiveness. This may be due to the following. First, some of the poorly translated phrases are not very important to the queries they appear in. *mejor artículo* means *best item*, but is translated as *best thing*. Second, at least one of the constituent term translations for each poorly translated phrase is correct. The effect of disambiguating at least one of the terms may reduce the overall negative effect of failing to translate the phrase. The phrase *prueba de inflación* meaning *inflation-proof* was translated as *inflation evidence*. In this case, the key term *inflación* was translated correctly. Table 4 gives the effect that translating phrases had on query effectiveness. It shows precision values for word-by-word with phrase dictionary translation (PD) versus word-by-word with co-occurrence translation (CO) and word-by-word with phrase dictionary and co-occurrence translation (PD+CO) as compared to the baseline of word-by-word (WBW) translation. Each of the queries containing correct CO phrasal translations improved. The improvement in effectiveness with the addition of CO over PD alone is significant at the .01 level. The addition of phrasal translations using both methods brings cross-language effectiveness up to 79% of mono-lingual as measured by average precision. In fact, only half of the queries in which phrases were translated via co-occurrence information do worse than their monolingual counterparts. Translation without phrases yields only 60% of monolingual.

Query	WBW	PD	CO	PD+PLC	PD+CO
Avg.Prec.	0.2331	.2944	0.2741	0.2551	0.3057
% change		26.3	17.6	9.4	31.1
Precision at:					
5 docs:	0.3619	0.3905	0.3714	0.4095	0.4190
10 docs:	0.3286	0.3714	0.3762	0.3857	0.4048
20 docs:	0.3095	0.3738	0.3690	0.3524	0.4048
30 docs:	0.2810	0.3413	0.3238	0.3254	0.3746

Table 4: Average precision for word-by-word translations and word-by-word translations augmented by both phrasal translation methods.

It should be noted that poor translations can decrease effectiveness as shown in [BC97]. One way to reduce this problem, could be to include more query terms in the co-occurrence analysis. Including more terms would provide more context and may further disambiguate translations. In particular, the inclusion of additional terms having unambiguous translations themselves would provide an anchor point. This anchor point would help to establish the correct context for the disambiguation.

6 Comparing Co-occurrence and Parallel Corpus Methods for Term Disambiguation

Parallel corpora can be used to disambiguate term translations as described in section 2.1. We showed in the above section that co-occurrence statistics can be used to disambiguate terms as phrasal constituents. We now show that that it could also be used for general term disambiguation and compare it to the parallel corpus technique.

We translated our query set in the following way. Phrases were translated using the phrase dictionary. Terms were translated word-by-word and then disambiguated using the parallel corpus method. We looked at sixty terms disambiguated by the parallel corpus and investigated how well they could be disambiguated via co-occurrence. We used the same co-occurrence method that was used for disambiguating phrase translations. However, rather than require the term be a phrase constituent, we paired the term to be disambiguated with an anchor. In this investigation, an anchor is a query noun that has an unambiguous translation, a proper noun, or a phrase translation. The resulting translations were then evaluated by a human assessor. Our conjecture was that co-occurrence disambiguation would not do any worse than parallel corpus disambiguation. Table 5 shows the overlap of terms correctly and incorrectly disambiguated by each method.

	correctly disamb. via parallel corpus	incorrectly disamb. via parallel corpus
correctly disamb. via co-occurrence	36	11
incorrectly disamb. via co-occurrence	3	10

Table 5: Term disambiguation overlap.

A sign test at the .05 level shows that the co-occurrence method is significantly better at disambiguating than is the parallel corpus method. When the co-occurrence method does not correctly disambiguate a term, there appears to not be enough context to infer the correct translation. The translation of *Efectos del chocolate en la salud. Cuales, si existen, son los efectos del chocolate en la salud.* is *The effects of chocolate on health. What, if any, are the effects of chocolate on health?*. The Spanish word *chocolate* can be translated as *chocolate, cocoa,* or *blood*. Given that it is more common to find *blood* co-occurring with *health*, blood is chosen over the uncommon and correct translation *chocolate*. One means of ameliorating the problem could be through pre-translation expansion. This is described in more detail later, but the basic idea follows. Prior to translation, retrieval is performed with the source query on a source language database. The query is then expanded with the best terms from the top ranking passages retrieved in response to the query. These expansion terms may provide enough context to be good anchors for disambiguation. *Hershey*, a brand of chocolate, is one of the expansion terms for the example query given above. Using *Hershey* as an anchor, rather than one of the original query terms, will more likely disambiguate *chocolate* to *chocolate* than to *blood*.

The failure of the parallel corpus method to disambiguate seems to be related to there being few or no documents related to the query. This is a problem more likely to happen the narrower or the more different the domain of the parallel corpus is from the corpus being searched. Our experiments are based on the UN parallel corpus which contains documents concerned with international peace and security, and health and education in developing countries. The query set is more general. Although there will be some general vocabulary overlap, the lack of relevant documents may prevent the disambiguation of query specific concepts. The UN corpus does not, for example, contain any documents relating the effects of chocolate on health and the parallel corpus method incorrectly disambiguates *chocolate* to *blood*. Of course this remains conjecture and needs to be borne out experimentally. However, it suggests that the co-occurrence method will be a more effective disambiguation method than the parallel corpus technique. This may be especially true when we can not rely on domain specific resources or at least on there being more domain overlap.

Nearly all of the phrases not translatable via the phrase dictionary are translatable word-by-word. We were interested in comparing the effectiveness of parallel corpus disambiguation with co-occurrence disambiguation. Recall that for all queries, terms are translated word-by-word and noun phrases are translated via our phrase dictionary. The co-occurrence method (CO) disambiguates the remaining phrase term translations based on their co-occurrence with other terms in a phrase. The parallel corpus disambiguation method (PLC) uses query context to disambiguate all remaining terms whether or not they are constituents of a phrase. We also wanted to see how the PLC and CO methods compared to more sophisticated machine translation (MT) systems.

Using a baseline of word-by-word translation (WBW), table 6 compares the effectiveness of both PLC and CO with that of two MT systems. The first is a web accessible off-the-shelf package called T1 from Langenscheidt [GMS] and the second is the on-line SYSTRAN [Inc] system. This table also gives cross-language performance as a percentage of monolingual. The co-occurrence method is more effective and gives higher recall and higher precision at all recall levels than does the PLC method. The SYSTRAN MT system is about as effective as the PLC method. There is no significant difference between the Langenscheidt MT system and the CO method which attains 79% of monolingual effectiveness. This is encouraging because it shows that co-occurrence information can be successfully employed to attain the effectiveness of a reasonably effective MT system. This is a positive statement for the possibilities of cross-language searching in languages for which few resources exist or for which a reasonable MT system does not exist.

Method	Precision	%change	% Monolingual
Monolingual	0.3869		-
WBW	0.2331		60
PLC	0.2551	9.4	65
CO	0.3057	31.1	79
T1	0.3066	31.5	79
SYSTRAN	0.2584	10.8	67

Table 6: Average precision as a percentage of that for monolingual.

7 Combinations of Disambiguation Methods

Earlier work showed that query expansion can greatly reduce the error associated with dictionary translations. In the following experiments, we look at the effectiveness of combining the disambiguation methods described above with query expansion via Local Context Analysis (LCA). We first translated queries automatically via MRD as described in section 4. Phrases were translated using the phrase dictionary and then one of the corpus disambiguation methods described above was applied. The co-occurrence method was performed with a window size of 250 terms. Queries were then expanded via LCA prior to translation, after translation or both before and after translation. We also compared these results to the expansion of queries translated via the method reported in our earlier work [BC97] and which we refer to as "sense1".

The sense1 method proceeds as follows. Multi-term concepts are translated as phrases using the phrase dictionary. The remaining terms are translated word-by-word without the aid of part-of-speech. A dictionary entry may list several senses for a word, each having one or more translations. To reduce the number of extraneous terms, only the target translations corresponding to the first sense listed in the dictionary entry are taken. We assume that the first sense listed is also the most frequent. Finally, we use the #synonym operator to disambiguate a term translation containing more than one target equivalent. We did not do this in work reported previously, but do it here for consistency of comparison to the experiments in this study.

7.1 Pre-translation Expansion

The following set of experiments show how effective pre-translation expansion is for further disambiguating three types of query translations: the sense1 method, the parallel corpus disambiguation method (PLC), and the co-occurrence method (CO). Pre-translation expansion is done in the following way. The top 20 passages are retrieved in response to the source query. The query is then expanded with the top 5 source terms. Expansion is followed by query translation. Average precision values are given in table 7. Word-by-word translation as described in section 4 is used as a baseline. Columns two,four, and six are queries translated via the sense1, PLC, and CO methods, respectively. Columns three, five, and seven are the sense1, PLC, and CO methods each with pre-translation expansion. Earlier work showed that pre-translation expansion enhances precision. Results are consistent with this, with the exception of pre-translation expansion of the PLC disambiguated queries. The problem here is that many of the expansion terms were disambiguated incorrectly, so that nearly half of the queries lost effectiveness. The improvement in average precision of expanded co-occurrence disambiguated queries over co-occurrence disambiguation alone is not significant. This may be due to the improved quality of CO translation over the other translation methods. In other words, the CO method alone may be reducing much of the ambiguity that is reduced by pre-translation

Query	WBW	1st	1st+Pre	PLC	PLC+Pre	Co	Co+Pre
Avg.Prec.	0.2331	0.2392	0.2568	0.2551	0.2155	0.3057	0.3098
% change		2.6	10.1	9.4	-7.6	31.1	32.9
Precision at:							
5 docs:	0.3619	0.3238	0.3429	0.4095	0.3333	0.4190	0.4667
10 docs:	0.3286	0.2810	0.3190	0.3857	0.3476	0.4048	0.4333
20 docs:	0.3095	0.3119	0.2952	0.3524	0.3143	0.4048	0.3976
30 docs:	0.2810	0.2651	0.2714	0.3254	0.2857	0.3746	0.3683
100 docs:	0.1705	0.1676	0.1795	0.1929	0.1652	0.2443	0.2324

Table 7: Average precision and precision at low recall for word-by-word, sense1, sense1 with pre-translation expansion, parallel corpus disambiguation, parallel corpus disambiguation with pre-translation expansion, co-occurrence disambiguation, and co-occurrence disambiguation with pre-translation expansion.

expansion with other methods of translation.

7.2 Post-translation Expansion

In these experiments, post-translation LCA expansion was performed by addition of the top 50 concepts from the top 30 passages after query translation. All multi-term concepts were wrapped in INQUERY #PASSAGE25#PHRASE operators. Terms within this operator were evaluated to determine whether they co-occur frequently. If they do, the terms must be found within 3 words of each other to contribute to the document's belief value. If they do not co-occur frequently, the terms in the phrase are treated as having equal influence, however they must be found within twenty-five words of each other. Concepts were weighted with an Infinder-like [JC94] weighting scheme. The top ranked concept was given a weight of 1.0 with all subsequent concepts down-weighted by $\frac{T-i-1}{T}$, where T is the total number of concepts and i is the rank of the current concept. This weighting scheme was shown to be effective in LCA experiments for the TREC evaluations [VH96]. Expansion was carried out after translation of queries via either the sense1, PLC, or CO methods.

Table 8 shows average precision values for seven query sets. As in the previous section, Word-by-word translation is used as a baseline. Columns three, five, and seven are the sense1, PLC, and CO methods each with post-translation expansion. Our earlier work showed that post-translation expansion enhances recall and precision. These results are consistent with those findings. The most effective queries are those translated via CO followed by post-translation expansion. Recall is also higher for this query set.

7.3 Combined Pre- and Post-translation Expansion

The combination experiments start with the pre-translation LCA expansion of the source queries. After the expanded queries were translated automatically via the sense1, PLC, or CO method, they were expanded again via LCA multi-term expansion. The pre- and post- translation phases proceed as described in sections 7.1 and 7.2. Results are given in table 9.

As expected, combining pre- and post-translation expansion boosts both precision and recall. There is no significant difference between post-translation and combined expansion of the CO translated queries. This makes sense in light of the fact that the CO method appears to disambiguate queries so well that pre-translation expansion has little impact on effectiveness. There is no significant difference between CO expanded via the post-translation method or CO expanded via the combined method. However, the combined expansion method may be preferred here since precision is slightly higher at low recall.

Table 10 shows the effectiveness of each of the best expansion methods as a percentage of monolingual performance as measured by average precision. Results show that combining our disambiguation methods brings cross-language performance to more than 90% of monolingual performance.

8 Conclusions and Future Work

One of the main hurdles to improving cross-language retrieval effectiveness has been the reduction of ambiguity associated with query translation. Translation error is due largely to addition of extraneous terms and failure to correctly translate phrases. In addition, the resources needed to address this problem typically require considerable manual effort to construct and may be difficult to acquire.

A few simple techniques such as part-of-speech tagging and the use of the #synonym operator can address the extraneous term problem. Phrasal translation is more problematic. Certain types of multi-term concepts, such as proper noun phrases, are easily translated via MRD. However, dictionaries do not provide enough context for accurate phrasal translation in other cases. The correct translations of phrase terms tend to co-occur and incorrect translations tend not to co-occur. Corpus analysis can exploit this information to significantly reduce ambiguity of phrasal translations. Combining phrase translation via phrase dictionary and co-occurrence disambiguation brings CLIR performance up to 79% of monolingual. The co-occurrence technique can also be used to reduce ambiguity of term translations.

Query expansion via Local Context Analysis can be used to further reduce the error associated with query translation. Pre-translation expansion becomes less effective as query disambiguation improves. However, we believe pre-translation expansion terms may still be useful as anchors for disambiguation via the co-occurrence method. Post-translation expansion and combining pre- and post-translation expansion enhance both recall and precision. Combining either of these two expansion methods with query translation augmented by phrasal translation and co-occurrence disambiguation brings CLIR performance above 90% monolingual. Even with a higher baseline of monolingual with expansion, combining the CO method with expansion can still yield up to 88% of monolingual performance. This is a considerable improvement over previous work which yielded 68% monolingual.

In this study, we have shown that combining corpus analysis techniques can be used to disambiguate terms and phrases. In combination with query expansion, it significantly reduces the error associated with query translation. Techniques based on unlinked corpora can perform as well or better than techniques based on more complex or scarce resources. Our co-occurrence method was better at disambiguating queries than was our parallel corpus technique. In addition, it performed as well as a reasonable MT system. This suggests that we can effectively use readily available resources such as unlinked corpora to increase cross-language effectiveness. This will have an even larger im-

Query	WBW	1st	1st+Post	PLC	PLC+Post	Co	Co+Post
Avg.Prec.	0.2331	0.2392	0.3317	0.2551	0.2864	0.3057	0.3623
% change		2.6	42.3	9.4	22.8	31.1	55.4
Precision at:							
5 docs:	0.3619	0.3238	0.4476	0.4095	0.4000	0.4190	0.4857
10 docs:	0.3286	0.2810	0.4333	0.3857	0.3857	0.4048	0.4857
20 docs:	0.3095	0.3119	0.3905	0.3524	0.3667	0.4048	0.4429
30 docs:	0.2810	0.2651	0.3651	0.3254	0.3476	0.3746	0.4111
100 docs:	0.1705	0.1676	0.2452	0.1929	0.2167	0.2443	0.2838

Table 8: Average precision and precision at low recall for word-by-word, sense1, sense1 with post-translation expansion, parallel corpus disambiguation, parallel corpus disambiguation with post-translation expansion, co-occurrence disambiguation, and co-occurrence disambiguation with post-translation expansion.

Query	WBW	1st	1st+Comb	PLC	PLC+Comb	Co	Co+Comb
Avg.Prec.	0.2331	0.2392	0.3193	0.2551	0.2593	0.3057	0.3533
% change		2.6	37.0	9.4	11.2	31.1	51.5
Precision at:							
5 docs:	0.3619	0.3238	0.3905	0.4095	0.3619	0.4190	0.4952
10 docs:	0.3286	0.2810	0.4190	0.3857	0.3333	0.4048	0.4810
20 docs:	0.3095	0.3119	0.4024	0.3524	0.3357	0.4048	0.4452
30 docs:	0.2810	0.2651	0.3556	0.3254	0.3095	0.3746	0.3968
100 docs:	0.1705	0.1676	0.2424	0.1929	0.2019	0.2443	0.2690

Table 9: Average precision and precision at low recall for word-by-word, sense1, sense1 with post-translation expansion, parallel corpus disambiguation, parallel corpus disambiguation with post-translation expansion, co-occurrence disambiguation, and co-occurrence disambiguation with post-translation expansion.

pact on cross-language retrieval between languages for which relatively few resources exist.

Method	Precision	% Monolingual
Mono	0.3869	-
CO+pre	0.3098	80
sense1+post	0.3317	86
CO+post	0.3623	94
CO+combined	0.3533	91

Table 10: Average precision as a percentage of that for monolingual.

Acknowledgments

This material is based on work supported by the National Science Foundation, Library of Congress, and Department of Commerce under cooperative agreement number EEC-9209623.

References

[AF77] R. Attar and A. S. Fraenkel. Local feedback in full-text retrieval systems. *Journal of the Association for Computing Machinery*, 24:397–417, 1977.

[BC96] Lisa Ballesteros and W. Bruce Croft. Dictionary-based methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, pages 791–801, 1996.

[BC97] Lisa Ballesteros and W. Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th International Conference on Research and Development in Information Retrieval*, pages 84–91, 1997.

[CCB95] J.P. Callan, W.B. Croft, and J. Broglio. Trec and tipster experiments with inquery. *Information Processing and Management*, 31(3):327–343, 1995.

[CYF^{+}97] J. G. Carbonell, Y. Yang, R. E. Frederking, R. Brown, Y. Geng, and D. Lee. Translingual information retrieval: a comparative evaluation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'97)*, 1997.

[Dav96] Mark Davis. New experiments in cross-language text retrieval at nmsu's computing research lab. In *Proceedings of the Fifth Retrieval Conference (TREC-5) Gaithersburg, MD: National Institute of Standards and Technology*, 1996.

[DIS91] Ido Dagan, Alon Itai, and Ulrike Schwall. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 130–137, 1991.

[DO97] Mark W. Davis and William C. Ogden. Quilt: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th International Conference on Research and Development in Information Retrieval*, pages 92–98, 1997.

[GLY96] Denis A. Gachot, Elke Lange, and Jin Yang. An application of machine translation technology in multilingual information retrieval. In *Working notes of the Workshop on Cross-linguistic Information Retrieval*, pages 44–54, 1996.

[GMS] Gesellschaft fuer multilinguale Systeme GMS. http://www.gmsmuc.de/english/ (Jan. 1998).

[Har97] Donna Harman, editor. Proceedings of the 6th Text Retrieval Conference (TREC-6). 1997.

[HG96] David A. Hull and Gregory Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pages 49–57, 1996.

[Inc] SYSTRAN Software Inc. http://www.systranet.com (Jan. 1998).

[JC94] Y. Jing and W.B. Croft. An association thesaurus for information retrieval. In *RIAO 94 Conference Proceedings*, pages 146–160, 1994.

[KH97] Wessel Kraaij and Djoerd Hiemstra. In *To appear in Proceedings of the Sixth Retrieval Conference (TREC-6) Gaithersburg, MD: National Institute of Standards and Technology*, 1997.

[Kup93] Julian M. Kupiec. An algorithm for finding noun phrase correspondances in bilingual corpora. In *Proceedings, 31st Annual Meeting of the ACL*, pages 17–22, 1993.

[LL90] Thomas K. Landauer and Michael L. Littman. Fully automatic cross-language document retrieval. In *Proceedings of the Sixth Conference on Electronic Text Research*, pages 31–38, 1990.

[SB96] Paraic Sheridan and Jean Paul Ballerini. Experiments in multilingual information retrieval using the spider system. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pages 58–65, 1996.

[SBS97] Paraic Sheridan, Martin Braschler, and Peter Schauble. Cross-language information retrieval in a multilingual legal domain. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, pages 253–268, 1997.

[SMH96] Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, 1996.

[TC91a] Howard R. Turtle and W. Bruce Croft. Efficient probabilistic inference for text retrieval. In *RIAO 3 Conference Proceedings*, pages 664–661, 1991.

[TC91b] Howard R. Turtle and W. Bruce Croft. Inference networks for document retrieval. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pages 1–24, 1991.

[VH96] E.M. Voorhees and D.K. Harman, editors. Proceedings of the 5th Text Retrieval Conference (TREC-5). 1996.

[vR77] C. J. van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33:106–119, 1977.

[XC96] Jinxi Xu and W. Bruce Croft. Querying expansion using local and global document analysis. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pages 4–11, 1996.

[XC98] Jinxi Xu and W. Bruce Croft. Corpus-based stemming using co-occurrence of word variants. *To appear in ACM TOIS*, January, 1998. Technical Report TR96-67, Dept. of Computer Science, University of Massachusetts/Amherst.