Draft copy: Please do not distribute. IR-99

Statistical Methods for Cross-Language Information Retrieval

Lisa Ballesteros and W. Bruce Croft
balleste@cs.umass.edu, croft@cs.umass.edu
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts
Amherst, MA 01003-4610 USA

Abstract

Multi-lingual information retrieval (IR) has largely been limited to the development of multiple systems for use with a specific foreign language. The explosion in the availability of electronic media in languages other than English makes the development of IR systems that can cross language boundaries increasingly important. We are currently developing tools and techniques for Cross Language Information Retrieval. In this paper, we present experiments that analyze the factors that affect dictionary based methods for cross-language retrieval and present methods that dramatically reduce the errors such an approach usually makes.

1 Introduction

In recent years, the amount of online information from the government, scientific, and business communities has risen dramatically. In response to this increase, much work has been done to develop mono-lingual systems that provide effective and efficient access to electronic media. However, the diversity of information sources and the explosive growth of the Internet worldwide are compelling evidence of the need for IR systems that can cross language boundaries. Increased interchange within the international community would be greatly facilitated by multi-lingual Information Retrieval (IR) techniques.

Machine translation is a growing area of research that could address some of the issues of multiple language environments. However, the necessary linguistic analysis is currently expensive to implement and its computational complexity can be prohibitive. In addition, linguistic techniques alone do not address issues of access and retrieval, and the full translation of a document may be unnecessary for the assessment of a document's relevance.

Our goal for multi-lingual IR is to enable a user to query in one language but perform retrieval across languages. The term multi-lingual has also been associated with modifying systems to run in several mono-language retrieval modes. To differentiate between these two definitions the former is referred to as *cross-language information retrieval* (CLIR). CLIR would be useful for people who do not speak a foreign language well, but who can read it well enough to judge a document's relevance. A second advantage of this type of retrieval is that it would help to reduce the number of irrelevant documents that would otherwise have to be translated manually.

In this paper, we discuss a particular approach to CLIR based on bilingual dictionaries. We will show how queries can be "translated" using dictionaries, and we apply a process called "local feedback" to dramatically reduce the errors such an approach normally makes. Previous work in multi-lingual and cross-language IR is discussed in section 2. Section 3 discusses dictionary methods. We present our experimental methods and results in section 4 and discuss conclusions and future work in section 5.

2 Previous Work

Effective systems for mono-lingual information retrieval have been available for several years. However, despite the fact that cross-language retrieval would be useful in many settings, most of the research in this area has focused on incorporating new languages into existing systems. The modified systems can then run in several mono-language retrieval modes.

Salton [SB90] and Pevzner [Pev72] showed early on that with carefully constructed thesauri, cross-language retrieval was nearly as effective as mono-lingual retrieval. Salton's approach was to manually assign thesaurus classes to the terms contained in a small collection of French documents and their translations. Groups of related words from each language were placed in individual classes in such a way that corresponding groups from both languages were assigned the same class identifier. The thesaurus classes then acted as an interlingua between the two sets of documents. Term weighting was also used so that terms judged to be better discriminators were assigned higher weights. Results were nearly as good as the mono-lingual baseline, although, cross-language retrieval was less effective for French queries and English documents. The main problem was the variability of French which caused the translations of English terms from a single class to be mapped to French terms spanning several other classes. Thus some terms had no correct mapping from French to English and vice-versa. For example, assume English term $word_{e1}$ from $class_{10}$ has French

translations $word_{f1}$ and $word_{f2}$ from $class_{10}$ and $class_{23}$, respectively. $word_{f2}$ in a French query would be replaced with English terms from $class_{23}$, missing the correct translation in $class_{10}$. These studies were promising however, the test collection was very small by current standards and it is unrealistic to manually index today's larger databases. Also, no system was ever implemented so it is unclear how such a system would perform in practice.

Landauer and Littman [LL90] have also proposed a method for cross-language retrieval. Latent Semantic Indexing (LSI) [FDD+88] was used to create a multidimensional indexing space for a parallel corpus of English documents and their French translations. Their method has been successful at the task of retrieving a query's translation, in response to that query. However the collection used was small, containing 2482 paragraph-length documents from Canadian Parliamentary proceedings and no results of its effectiveness on the traditional retrieval task have been reported. The method also relies on the use of parallel corpora which are not always available.

Another method that relies on parallel and aligned corpora has been suggested by Dunning and Davis [DD93]. Their method is based on the vector space model and involves the linear transformation of the representation of a query in one language to its corresponding representation in another language. The transformation is done by reduction of the document space to generate a translation matrix. They have had some success in efficiently estimating the translation matrix and results of tests to estimate its quality are promising. However further tests of the effectiveness of the method have been limited by its computational complexity.

More recently, Davis and Dunning[DD95a, DD95b] have developed several other approaches to query translation, which they tested on the TREC ISM Spanish queries and collection. Two of these rely on the use of a Spanish-English parallel corpus and one uses evolutionary programming for query optimization. The first of the parallel corpus approaches extracts high frequency terms from the top 100 documents retrieved in response to a query. English queries were translated by replacing the original query terms with the 100 most frequent terms in the top 100 retrieved documents from the Spanish side of the parallel corpus. The second approach begins by extracting the terms from the top 100 retrieved sentences from the Spanish side of the parallel corpus. Terms found to be statistically significant were used to replace the original query terms. The Evolutionary programming method starts with a query generated by the high frequency approach. It then modifies queries by randomly adding or deleting query terms. Optimization is done by evaluating query fitness after each round of mutations, and selecting the "most fit" to continue to the next generation. Word-by-word translation was chosen as a baseline. The Evolutionary programming approach was the most effective, but results were disappointing, with each of the methods performing well below the baseline.

More recently, Sheridan and Ballerini [SB96] utilized co-occurrence thesauri generated from a comparable corpus. Comparable corpora are generated from collections of texts in pairs or multiples of languages. Documents that are about the same event or convey the same information are concatenated to create multi-lingual documents. These documents comprise the comparable corpus. Cross-language experiments suggest that using co-occurrence thesauri generated with this type of data yields a translation effect. In other words, words co-occurring in the multi-lingual documents are rough translations of each other. However, performance measured by average precision is still considerably below that of mono-lingual retrieval. Disadvantages to the approach are that it relies on time-sensitive documents, queries are constrained to referencing specific events, and a strict definition of the notion of relevance. Also, like parallel corpora, comparable corpora are not readily available.

3 Dictionary Translation Using Expansion

Although cross-language retrieval methods based on the use of parallel and aligned corpora have shown promise, one disadvantage is lack of resources. Parallel corpora are not always readily available and those that are available tend to be relatively small or to cover only a small number of subjects. Performance is also dependent on how well the corpora are aligned.

We are interested in finding methods for performing cross-language retrieval which do not rely on scarce resources. Bilingual machine-readable dictionaries (MRDs) which are more prevalent than parallel texts seem to be a good alternative. The coverage of MRDs, while not deep, is broad enough to be used for translations of queries covering a wide variety of topics. However, simple translations tend to be ambiguous and give poor results. Our experiments have shown a drop of about 50% in average precision when MRD translated queries are compared to mono-lingual performance [BC96]. These results support work done by Hull and Grefenstett [HG96]. Our main hypothesis is that query expansion via "local feedback" will improve the retrieval effectiveness of simple dictionary translation.

Relevance feedback [SB90] is a method by which a query is modified using information derived from documents whose relevance to the query is known. Typically, terms found in relevant documents are added to the query. Local feedback [AF77] differs from classic relevance feedback in that it assumes the top retrieved documents are relevant. Applying feedback prior to translation (pre-translation) should create a stronger base query for translation by adding terms that emphasize query concepts. Feedback after translation (post-translation) should reduce the effects of irrelevant query terms by adding more context specific terms. Dictionary definitions tend to give several senses each having one or more related meanings. Irrelevant terms are added when the translation of a word consists of a group of related terms that are used in different contexts. For example, the Spanish translation of "retard" could be either "retardar" or "retrasar" both meaning "to slow", however the former is used in the context of growth and the latter is used in the context of progress. The next section describes our experiments and the results.

4 Experiments

To make these studies more tractable, we limited ourselves to two languages: Spanish and English. The English queries consisted of the description fields of TREC [Har96] topics 151-171 and averaged 10.6 terms per query. The Spanish queries consisted of TREC topics SP26-45 and averaged 4.3 terms per query. Evaluation was performed on the English documents contained in the 2 GB Tipster (vol. 2) collection and the 208 MB TREC ISM (El Norte) Spanish collection

using relevance judgments for each mono-language query-set/database pair. Training data for the pre-translation feedback experiments was the ISM collection described above and the 301 MB San Jose Mercury News (SJMN) database from the Tipster collection.

Each English query has a set of relevance judgments or English documents which were pre-judged to be relevant to the query. In order to use these judgments, we needed to test the effectiveness of MRD translations to English. To do this, we created base queries by manually translating the English queries to Spanish (herein referred to as ES-BASE). The MRD translations of the base queries could then be evaluated using the relevance judgments of the original queries. Spanish queries were treated similarly (Spanish base queries are referred to as SE-BASE), and were evaluated using relevance judgments for Spanish. The manual translation of the Spanish queries was performed by a bilingual graduate student whose native language is English. The manual translation of the English queries was performed by a bilingual graduate student whose native language is Spanish.

MRD translations were performed after simple morphological processing of query terms to remove most plural word forms and to replace some Spanish verb forms with their infinitive form. Translation is not used here in the sense of deep linguistic analysis. The terms of a query in the source language were merely replaced with the dictionary definition of those terms in the target language. Stop words and stop phrases such as "A relevant document will" were also removed. Dictionary definitions tend to give several senses each having one or more related meanings. To reduce ambiguity, we chose to replace query words with only those meanings given for the first sense of the definition. Table 1 gives some examples of the dictionary entries for the first sense of several words. The negative effect of this is that some relevant meanings will be lost. Words which were not found in the dictionary were added to the new query without translation. The Collins English-Spanish and Spanish-English bilingual MRDs were used for the translations.

term	meaning	MRD translation
mundo	world	people society secular life
conoce	know	know to know about understand meet get to know to become acquainted with
country	país	país patria campo región tierra

Table 1: Examples of terms, their meanings in particular queries, and their MRD word-by-word translation.

11-point average precision is used as the basis of evaluation for all experiments. It is unrealistic to expect the user to read many retrieved foreign documents to judge their relevance, so we also report precision at low recall levels. The following sections describe our experiments. First we analyzed the factors affecting word by word translation. We then applied local feedback techniques before or after MRD translation and describe how each method helps to improve performance. Finally, we combined pre-translation and post-translation feedback and discuss its effectiveness. Fig. 1 is a flow chart of query processing for the feedback experiments. 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 elsewhere[TC91b, TC91a, BCC94].

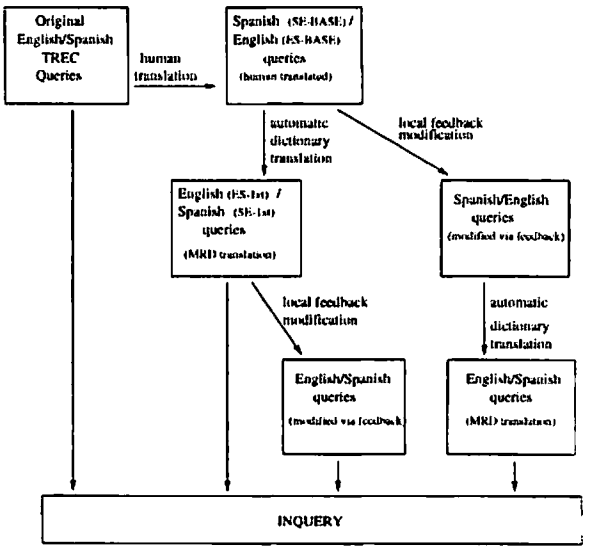

Figure 1: Flow chart of query processing.

4.1 Simple Word-By-Word Translation

Our first experiment was designed to test the effects of simple word-by-word MRD translation on retrieval performance and to determine what factors cause them. Queries manually translated from English and from Spanish were re-translated automatically word-by-word as described above. Briefly, each query term was replaced by the word or group of words given for the first sense of the term's definition. We refer to the re-translations of the English and Spanish base query sets as ES-1st and SE-1st respectively.

The re-translated queries lead to a 50-60% drop in performance as measured by average precision. We noted that these new queries were ambiguous in that they contained many more than one translation for some terms. Recall that the first sense of a dictionary definition may contain one or more related words. In addition to this, some query terms are more accurately translated via the translation of a phrase as shown in Table 2.

Phrase	meaning	MRD translation
cifras del costo	cost figures	amount of cost
fondos de inversión	mutual funds	fund of investment
marina de guerra	navy	navy of war
tratado de libre comercio	NAFTA	treaty of free trade

Table 2: Examples of phrases, their meanings and their word-by-word translations.

To try to determine the extent to which each of these factors was responsible for performance drops, two additional translations of the base queries were generated by hand. The first was a word by word translation in which we chose the one best term to replace each base query term (Manual-WBW). The second was like the first but with phrasal translation where appropriate (WBW+Phrases). The results given for English in Figure 2 show that performance does improve with the refinement of each query set. For comparison, these figures include performance of the original, untranslated queries. Results for Spanish queries are similar. The transfer of senses inappropriate to the query accounts for

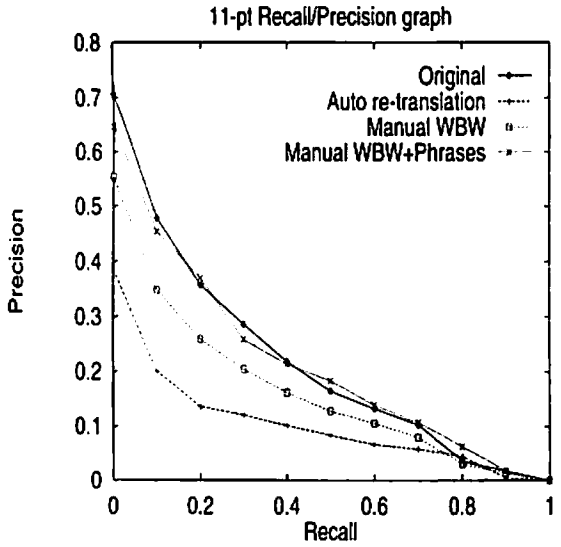

Figure 2: Query performance for the original English queries and three different methods of translation.

	query-terms	undefined translation	terms per
Original Span.	4.35	N/A	N/A
SE-BASE	4.95	12	N/A
SE-1st	17.6	N/A	4.05
Original Eng.	10.6	N/A	N/A
ES-BASE	10.75	5	N/A
ES-1st	32.45	N/A	3.09

Table 3: Query term statistics after removing stopwords.

12-29% of the loss of effectiveness and is a bigger problem for longer queries (English). Phrase loss accounts for 20-25%. An additional 12.4% of the loss from SE-BASE can be attributed to the exclusion from manual translations of acronyms that were in the original queries. This could also be expected to occur when queries contain technical terms or specialized vocabulary that are not found in general dictionaries. The remaining loss can probably be attributed to less well specified queries and to ambiguity introduced through manual translation and re-translation. Table 3 gives term statistics for the original queries, their manual translations (BASE) and their re-translations. The first column is query set name, second is the average number of terms per query, third is the number of terms in the set that were not found in the MRD, and last is the average number of terms returned from the MRD per translated query term.

In the following experiments, we tested the effectiveness of local feedback for reducing ambiguity. We applied local feedback to each query set either prior to (pre-translation) or after (post-translation) MRD translation. Pre-translation feedback modification should improve results by adding terms that emphasize query concepts. Post-translation feedback is expected to decrease ambiguity by de-emphasizing inappropriate terms.

4.2 Pre-translation Query Modification

We expected pre-translation feedback to be more effective with Spanish base queries since they are shorter than the English base queries. Fewer query terms means fewer content bearing terms, yielding a translation that is swamped by irrelevant words. Tables 4 and 5 show how ambiguity can reduce query effectiveness and how pre-translation feedback can reduce that ambiguity. In each example there are five representations of the same query: the original TREC query, the manual translation (BASE), the MRD translation of BASE, BASE after pre-translation feedback, and the MRD translation of those modified by pre-translation feedback. Words in parentheses were returned as a multiple word translation for one term in the preceding representation. Terms in brackets were added by feedback.

relaciones económicas y comerciales de México con los paises asiáticos , por ejemplo Japon, China y Corea
economic and commercial relations between Mexico and the Asiatic countries, for example, Japan, China, and Korea .
(económico equitativo rentable)comercial(narración relato relación)(Méjico México)(asiático)(país patria campo región tierra) el Japón China Corea
economic commercial relations mexico japan china korea korean nuclear south north]
(económico equitativo rentable)(comercial)(narración relato relación)(mexico)(laca japonesa)(porcelana loza)(korea) (korean)(nuclear)(sur mediodía)(norte)

Table 4: Five query representations for SP28: original, base, mrd translation of base, pre-translation feedback of base, and MRD translation of the pre-translation query.

programas para reprimir o limitar epidemias en México
programs for suppressing or limiting epidemics in Mexico
(programs)(controlador mayoritario)(restrictivo)(epidémico) (Méjico México)
programs controlling limiting epidemics mexico [epidemic cholera disease health]
(programs)(controlador mayoritario)(restrictivo)(epidémico) (mexico)(epidémico)(cólera)(enfermedad morbo dolencia mal) (salud sanidad higiene)

Table 5: Five query representations for SP43: original, base, MRD translation of base, pre-translation feedback of base, and MRD translation of pre-trans.

Performance of query SP28 gets worse with each translation. The problem with the first translation is that although all the original query terms are included, the query seems to get swamped by inappropriate word definitions. This is also a problem with pre-translation feedback. This problem is exacerbated because feedback returns all lowercased terms which is an artifact of tokenizing/indexing. Consequently, dictionary lookup fails to find proper nouns or instead finds their common noun definition: e.g., china is translated to "porcelana, loza" which mean "porcelain" and "china plate" respectively. This latter error can be minimized by ensuring that proper nouns retain capitalization. Note that this is less of a problem when translating from Spanish to English since fewer proper nouns are capitalized: e.g., the translation of Australian is australiano.

The word by word translation of query SP43 also suffers from the same problem described above. However, the pre-translation feedback improves performance considerably. The reason for this is that the inclusion of feedback terms related to epidemics and epidemic control strengthened the base query thus reducing the ambiguity of the translation.

The results in Tables 6 and 7 show that performance of SE-BASE and ES-BASE improves by up to 34% and 16%, respectively. In both cases, pre-translation feedback modification improves precision. The best results for ES-BASE resulted with the addition of 20 feedback terms. The improvement was limited by the increase in inappropriate translation terms.

Fdbk Terms	0	20	10
Train. Docs	0	10	10
Average Precision:			
Avg	0.0922	0.1072	0.0961
% Change:		16.4	4.3
Precision:			
5 docs:	0.2100	0.2300	0.2600
10 docs:	0.2050	0.2250	0.2300
15 docs:	0.2000	0.2233	0.2167
20 docs:	0.1900	0.2050	0.2075
30 docs:	0.1717	0.2050	0.1950

Table 6: Best pre-translation feedback results for ES-BASE queries.

Fdbk Terms	0	5	5
Train. Docs	0	30	50
Average Precision:			
Avg	0.0823	0.1099	0.1021
% Change:		33.5	24.0
Precision:			
5 docs:	0.2000	0.2500	0.2600
10 docs:	0.2100	0.2300	0.2600
15 docs:	0.1867	0.2400	0.2433
20 docs:	0.1975	0.2375	0.2350
30 docs:	0.1900	0.2217	0.2217

Table 7: Best pre-translation feedback results for SE-BASE queries.

4.3 Post-translation Query Modification

Post-translation feedback modification was expected to be more effective for ES-BASE than was pre-translation modification. The ES-BASE queries are longer than the SE-BASE queries thus tended to provide a strong base query for translation. Feedback should add more good terms which would help to reduce the affect of inappropriate translation terms.

Experimental results are given in Tables 8 and 9. Post-translation modification tends to improve recall with ES-BASE and SE-BASE queries showing improvements of up to 47.5 and 14.3%, respectively. Long queries show greater improvements. They are better specified so their MRD translations provide more context for improvement via feedback expansion. Table 10 illustrates the improvement typically gained by long queries. It asks about results stemming from the use by religious groups of the political process to further their goals and ambiguity is introduced after automatic re-translation. The query is refocused on its original intent by post-translation feedback via the addition of terms related to religion.

Fdbk Terms	0	5	5	5
Train. Docs	0	10	30	50
Average Precision:				
Avg	0.0922	0.1252	0.1346	0.1359
% Change:		35.8	46.1	47.5
Precision:				
5 docs:	0.2100	0.2600	0.2400	0.2400
10 docs:	0.2050	0.2300	0.2300	0.2350
15 docs:	0.2000	0.1967	0.2267	0.2300
20 docs:	0.1900	0.1875	0.2125	0.2200
30 docs:	0.1717	0.1750	0.1950	0.2000

Table 8: Best ES-BASE post-translation feedback results.

Fdbk Terms	0	20	20	30
Train. Docs	0	10	50	10
Average Precision:				
Avg	0.0823	0.0910	0.0916	0.0913
% Change:		10.6	11.3	10.9
Precision:				
5 docs:	0.2000	0.2500	0.1800	0.2300
10 docs:	0.2100	0.1950	0.1850	0.1800
15 docs:	0.1867	0.1900	0.1800	0.1800
20 docs:	0.1975	0.1975	0.1575	0.1800
30 docs:	0.1900	0.1633	0.1483	0.1617

Table 9: Best SE-BASE post-translation feedback results.

Spanish query 28 did not show improvement when feedback terms were added prior to translation. However, feedback after translation improved performance by 47% over the base query formulation. This improvement is probably due in part to the reduction of ambiguity caused by the reduction in inappropriate definitions such as "porcelana" and "loza". The inclusion of several terms related to commerce also helps to reduce ambiguity by de-emphasizing outliers. Table 11 shows the differences between four representations of query SP28, each is stemmed. The first is the original TREC query, the second is the MRD translation of BASE (BASE is the second representation of Table 4), the third is the MRD translation of the pre-translation modified BASE query, the fourth is the MRD translation of BASE after pre-translation feedback, and the last is the post-translation modification of MRD tranlated BASE. Performance of query SP28 is not typical of the shorter SE-BASE queries; they tend to show little improvement. They are less well specified so their MRD translations tended to be more ambiguous and post-translation feedback did not reduce this ambiguity.

4.4 Combined Feedback

In these experiments, we combined pre- and post-translation in the following way: base queries were modified via feedback, the modified queries were translated via MRD, the translated queries were modified via feedback, and then the MRD translations of the latter query set were used for evaluation.

The document will analyze the implications of the decision by Christian fundamentalists to use the political process to promote their objectives
El documento analizará las implicaciones de la decisión de los grupos fundamentalistas cristianos de derecha de usar el proceso político para promover su objetivo.
contradiction (decision judgment)(group cluster clump group) fundamentalist Christian (right hand right side right-hand side)(use make use wear) process political (promote advance further promote pioneer sponsor begin set on foot get moving bring) objective
contradiction decision judgment group cluster clump fundamentalist christian make wear process political promote advance pioneer sponsor begin set foot move bring objective religious religion church evangelist god]

Table 10: Four query representations of English query 155: original, base (manual translation), MRD re-translation of base, and post-translation feedback modification of the re-translated base query.

relacion econom comerc mex pais asiat japon chin cor
econom equit rentabl comerc narr relat rel mej mex asiat pai patri camp region tierr japon chin cor
(económico equitativo rentable) comercial (narración relato relación) mexico (laca japonesa)(porcelana loza) korea korean nuclear (sur mediodía) norte seoul soviet asia pyongyang japanese comunista (comercio negocio tráfico industria) asian (union enlace sindicato gremio obrero unión manguito unión) diplomático beijing unido península roh
econom equit rentabl comerc narr relat rel mej mex asiat pai patri camp region tierr japon chin cor [pais export asi comerci singapur kong merc taiw hong product japones industr invers canada millon dol malasi estadounidens tailandi import]

Table 11: Five query representations for SP28: original (stemmed), base(stemmed), MRD translation after pre-translation feedback, post-translation feedback (stemmed)

The combined method was most effective, yielding up to a 51% improvement in average precision. For the post-translation feedback stage with SE-BASE, eleven of the twenty base queries were improved by combined-feedback with 20 feedback terms from the top 50 documents. Of those eleven, six improved as a result of combining pre-translation and post-translation feedback, four others showed similar improvements to pre-translation alone, and 1 repaired the damage done by pre-translation feedback. The queries sets for other combined-feedback runs show similar results. As would be expected, both precision and recall are improved by the combined method. For ES-BASE, ten queries were improved by combined-feedback, but two of these did better with post-translation alone. An additional seven queries dropped below the improvement gained by post-translation alone. The remaining query showed no improvement over post-translation. Results are shown in Tables 12 and 13.

The improvements in performance occur because better query terms are added after the final feedback. Those terms tend to fine tune the query and de-emphasize inappropriate definitions and is illustrated in Figures 3 and 4. Combining

Fdbk Terms	0	10	20
Train. Docs	0	30	50
Average Precision:			
Avg	0.0823	0.1166	0.1242
% Change:		41.7	51.0
Precision:			
5 docs:	0.2000	0.2400	0.2600
10 docs:	0.2100	0.2200	0.2200
15 docs:	0.1867	0.2200	0.2000
20 docs:	0.1975	0.2075	0.2125
30 docs:	0.1900	0.1817	0.2017

Table 12: Best SE-BASE combined pre-translation and post-translation feedback results.

Fdbk Terms	0	5	20
Train. Docs	0	30	30
Average Precision:			
Avg	0.0922	0.1372	0.1375
% Change:		48.8	49.2
Precision:			
5 docs:	0.2100	0.2600	0.2700
10 docs:	0.2050	0.2450	0.2500
15 docs:	0.2000	0.2433	0.2467
20 docs:	0.1900	0.2400	0.2350
30 docs:	0.1717	0.2217	0.2300

Table 13: Best ES-BASE combined pre-translation and post-translation feedback results.

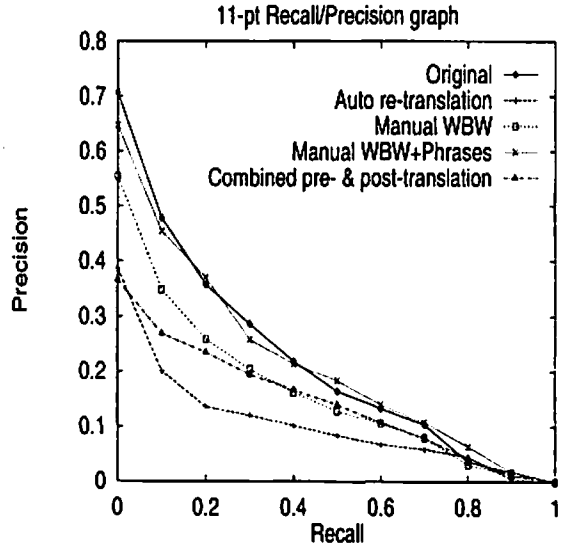

Figure 3: Query performance for five query sets: original English queries, MRD re-translation, manual word-by-word re-translation, manual word-by-word plus phrasal re-translation, combined pre- and post-translation feedback.

pre-translation and post-translation feedback reduces most of the error caused by the addition of extraneous terms via the translation process. For shorter queries, the method also reduces the error due to failure to accurately translate phrases and specialized vocabulary.

For both the ES-BASE and SE-BASE queries, those queries that showed improvement via pre-translation alone gained

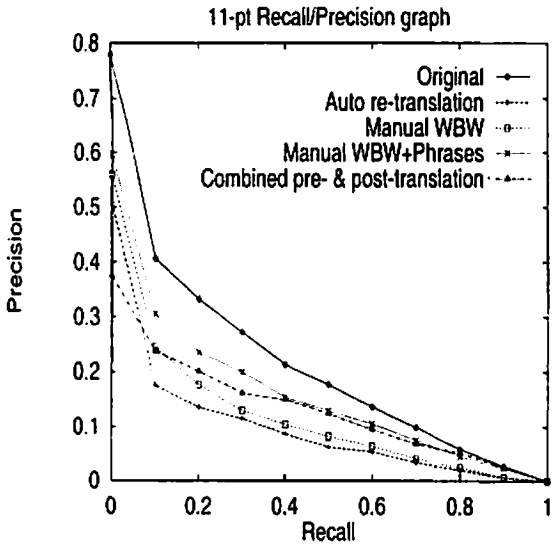

Figure 4: Query performance for five query sets: original Spanish queries, MRD re-translation, manual word-by-word re-translation, manual word-by-word plus phrasal re-translation, combined pre- and post-translation feedback.

greater improvements from subsequent post-translation feedback. This suggests that the pre-translation feedback stage creates a better base for translation and then the post-translation stage reduces the negative effects of ambiguity caused by inappropriate term definitions.

5 Conclusions and Future Work

Automatic MRD query translation leads to a more than 50% drop in retrieval effectiveness due to translation ambiguity. One of the two main factors affecting this is the lexical transfer of definitions containing too many extraneous terms. In our studies, this accounted for 12-29% of the loss in performance and is a greater problem with long queries. The second factor is phrase loss caused by word by word translation of concepts that are more appropriately translated as phrases. Phrase loss accounts for 20-25% of the loss in performance, but could be more of a problem in a language, such as French, that is richer in phrase construction. In addition to this, queries containing domain specific terminology which is not found in general dictionaries were shown to suffer an additional loss in performance.

We have found three means by which to dramatically reduce the loss in performance caused by-word-by word translation. The application of local feedback prior to translation creates a stronger base for translation and improves precision. Shorter, less well specified queries (SE-BASE) show the greatest improvements.

Local feedback query modification after MRD translation improves recall. Short queries tend neither to improve, since their translations are ambiguous giving little context for improvement; nor does performance degrade after modification. The improvements to long queries are due to the introduction of terms which de-emphasize irrelevant translations and thus reduce ambiguity.

Combining the two above methods is also effective, yielding improvements of 50% in average precision. Short queries gain the greatest improvements from this combination because they are less well specified. The first feedback stage strengthens the original query providing a better base for translation. The second feedback stage reduces ambiguity by de-emphasizing irrelevant terms added by translation. For short queries, this method also reduces the error associated with the failure to translate multi-term expressions as phrases.

The use of these methods does not correct entirely the loss in performance due to word-by-word MRD translation. They are however, simple to apply and do not rely on scarce resources such as parallel corpora. Results show that half of the loss in performance can be regained by combining local feedback with MRD translation. The improvement is a result of reducing the negative effects of two factors: extraneous definitions and failure to translate phrases. In addition, this approach is significantly better than similar approaches taken previously with realistic document collections.

Although performance is still a long way from "optimal", i.e. original query performance, we are encouraged by our results. We believe that further improvements in dictionary based methods are possible through the use of other statistical methods. Ambiguity arising from word-by-word translation has been reduced by this approach. However, ambiguity arising from the word-by-word translation of phrases is still a major factor in the loss of performance. We are currently investigating methods to address this problem. One means for doing so is to create a bi-lingual phrase dictionary from MRD information. We are also exploring ways in which INFINDER [JC94] and local context analysis can be used to reduce phrasal ambiguity.

Acknowledgments

An early version of this paper appeared in Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications (1996). This research is supported by a NASA GSRP grant, #NGT-70358 and 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.

[BCC94] J. Broglio, J. Callan, , and W.B. Croft. Inquery system overview. In *Proceedings of the TIPSTER Text Program (Phase I)*, pages 47–67, 1994.

[DD93] Ted Dunning and Mark Davis. Multi-lingual information retrieval. Technical report MCCS-93-252, Computing Research Laboratory, New Mexico State University, 1993.

[DD95a] Mark Davis and Ted Dunning. Query translation using evolutionary programming for multilingual information retrieval. In *Proceedings of the Fourth Annual Conference on Evolutionary Programming*, 1995.

[DD95b] Mark Davis and Ted Dunning. A trec evaluation of query translation methods for multi-lingual text retrieval. In *In Proceedings of the Fourth Retrieval Conference (TREC-4) Gaithersburg, MD: National Institute of Standards and Technology, Special Publication 500-215*, 1995.

[FDD+88] G.W. Furnas, S. Deerwester, S.T. Dumais, T.K. Landauer anmd R.A. Harshman, L.A. Streeter, and K.E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th International Conference on Research and Development in Information Retrieval*, pages 465–480, 1988.

[Har96] Donna Harman, editor. *Proceedings of the 4th Text Retrieval Conference (TREC-4)*. 1996.

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

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

[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.

[Pev72] B. Pevzner. Comparative evaluation of the operation of the russian and english variants of the pusto-nepusto-2 system. *Automatic Documentation and Mathematical Linguistics*, 6:71–74, 1972.

[SB90] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297, 1990.

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