Dictionary Methods for Cross-Lingual Information Retrieval

Lisa Ballesteros and Bruce Croft

Computer Science Department
University of Massachusetts
Amherst, MA 01003 USA
{balleste,croft}@cs.umass.edu

Abstract. Multi-lingual information retrieval (IR) has largely been limited to the development of 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. In this paper, we present experiments that analyze the factors that affect dictionary based methods for cross-lingual 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, much work has been done to develop 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 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 is also associated with modifying systems to run in several mono-language retrieval modes. We differentiate between these definitions by referring to the former as *cross-lingual 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 understand a document's contents and judge its relevance. A second advantage of this type of retrieval is that it would help to reduce the number of irrelevant documents for manual translation.

In this paper, we discuss a particular approach to CLIR based on bilingual dictionaries. We 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. Sect. 2 discusses dictionary methods. Experimental methods and results are presented in Sect. 3. Related work in multi-lingual and cross-lingual IR is discussed in Sect. 4 and we discuss conclusions 5.

2 Dictionary Translation Using Expansion

We are interested in finding methods for performing cross-lingual retrieval which do not rely on scarce resources such as parallel corpora. Bilingual, machine-readable dictionaries (MRDs), 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 main hypothesis is that query expansion via "local feedback" will improve the retrieval effectiveness of simple dictionary translation.

Relevance feedback [10] 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 [1] 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.

3 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 [7] topics 151-171 and the Spanish queries consisted of TREC topics SP26-45. Evaluation was performed on the English documents contained in the 2 GB TREC (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 consisted of the documents in the ISM collection described above and in the 301 MB San Jose Mercury News (SJMN) database from the TREC collection.

Each English query has "relevance judgments", or a set of 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). 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 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 one language are merely replaced with the dictionary definition of those terms in another 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. Table 1 gives some examples of the dictionary entries for the first sense of several words. To reduce ambiguity, we chose to replace query words with only those meanings given for the first sense of the definition. 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.

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

Term	Meaning	MRD Translation
mundo	world	people, society, secular, life
conocer	know	know, to know about, understand, meet, get to know, to become acquainted with
country	país	país, patria, campo, región, tierra

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 and 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. In this paper, query sets beginning ES- and SE- refer to sets resulting from modifications made to the ES-BASE and SE-BASE queries, respectively. 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[12, 11, 2].

3.1 Simple Word-By-Word Translation

Our first experiment was designed to test the effects of simple word-by-word translation on retrieval performance and to determine the factors causing them.

Base queries were translated word for word via MRD 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. The translations of the English and Spanish base query sets are ES-1st and SE-1st, respectively.

The translated queries lead to a 50-60% drop in performance as measured by average precision as shown in Table 3. We noted that these new queries were ambiguous, containing 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.

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

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

To find 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 (ES-WBW and SE-WBW). The second was like the first but with phrasal translation where appropriate (ES-Phr and SE-Phr). The results given in Table 3 show that performance does improve with the refinement of each query set. The transfer of senses inappropriate to the query accounts for 12-29% of the loss of effectiveness while phrase loss accounts for 20-25%. An additional 12.4% of the loss from the translation of SE-BASE can be attributed to the exclusion of acronyms that were in the original queries. The rest can probably be attributed to less well specified queries and to ambiguity introduced through the original manual translation. Table 4 gives term statistics for the base queries and their translations. The first column is query set name, second is the average number of terms per query, third is the number of terms that were not found in the MRD, and the last is the average number of terms returned from the MRD per translated query term.

3.2 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 could mean fewer content bearing terms which might yield a translation that is swamped by irrelevant words. Tables 5 and 6 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 first is

Table 3. Average precision for ES and SE queries. Columns correspond to the following query sets: 2,6 - original, untranslated TREC queries; 3, 7 - MRD translated base queries; 4,8 - manual word by word translations; 5, 9 - manual phrasal translations; 10 - manual phrasal translation plus acronyms.

	ES queries				SE queries				
	Orig.	ES-1st	ES-WBW	ES-Phr	Orig.	SE-1st	SE-WBW	SE-Phr	SE-Phr +acron.
	0.2076	0.0922	0.1517	0.2030	0.2016	0.0823	0.1064	0.1473	0.1722
% Change:		-55.6	-26.9	-2.2		-59.2	-47.2	-26.9	-14.6

Table 4. Query term statistics after removing stop-words.

Query-set	Query-terms	Undefined	Terms per translation
Original Spanish	4.35	N/A	N/A
SE-BASE	4.95	12	N/A
SE-1st	17.6	N/A	4.05
Original English	10.6	N/A	N/A
ES-BASE	10.75	5	N/A
ES-1st	32.45	N/A	3.09

the original TREC query, second is the manual translation, third is the MRD translation of BASE, fourth is BASE after pre-translation feedback, and the last is the MRD translation of the latter. Words in parentheses were returned as a multiple word translation for one term. Terms in brackets were added by feedback.

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 inclusion of feedback terms related to epidemics and epidemic control strengthened the base query thus reducing the ambiguity of the translation.

Table 5. Five query representations for SP28: original, base, MRD translation of base, pre-translation feedback of base, and MRD translation of the pre-translation query.

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

Table 6. Five query representations for SP43: original, base, MRD translation of base, pre-translation feedback of base, and MRD translation of pre-translation feedback query.

Original	programas para reprimir o limitar epidemias en México
BASE	programs for suppressing or limiting epidemics in Mexico
MRD-translated BASE	(programs)(controlador mayoritario)(restrictivo)(epidémico) (Méjico México)
BASE+ pre-translation feedback	programs controlling limiting epidemics mexico [epidemic cholera disease health]
MRD-translated Base + pre-translation feedback	(programs)(controlador mayoritario)(restrictivo)(epidémico) (mexico)(epidémico)(cólera)(enfermedad morbo dolencia mal) (salud sanidad higiene)

The results in Table 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 from the top 10 documents. The improvement was limited by the increase in inappropriate translation terms.

3.3 Post-translation Query Modification

Post-translation feedback was expected to be more effective than pre-translation feedback for ES-BASE. ES-BASE queries are longer than 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.

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

Table 7. Best pre-translation feedback results.

	ES-BASE			SE-BASE			
Fdbk Terms	0	20	10	0	5	5	5
Train. Docs	0	10	10	0	10	30	50
Average Precision:							
	0.0922	0.1072	0.0961	0.0823	0.1014	0.1099	0.1021
% Change:		16.4	4.3		23.2	33.5	24.0
Precision:							
5 docs:	0.2100	0.2300	0.2600	0.2000	0.2600	0.2500	0.2600
10 docs:	0.2050	0.2250	0.2300	0.2100	0.2500	0.2300	0.2600
15 docs:	0.2000	0.2233	0.2167	0.1867	0.2433	0.2400	0.2433
20 docs:	0.1900	0.2050	0.2075	0.1975	0.2300	0.2375	0.2350
30 docs:	0.1717	0.2050	0.1950	0.1900	0.2017	0.2217	0.2217

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 8 shows the differences between four representations of query SP28, all but the third is stemmed.

Table 8. Five stemmed query representations for SP28: original, MRD-translated base, MRD translation after pre-translation feedback (unstemmed), post-translation feedback.

Original	relacion econom comerc mex pais asiat japon chin cor
MRD-translated BASE	(econom equit rentabl) comerc (narr relat rel) (mej mex) asiat (pai patri camp region tierr) japon chin cor
MRD-translated BASE + pre-translation feedback	(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
post-translation feedback of MRD-translated BASE	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]

Experimental results are given in Table 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.

Some of the difference in the performance of the pre-translation and post-translation methods may be explained by the difference in the quality of the training data. The translated feedback query for SP28 returned 11 relevant documents in the top 20 retrieved (used for post-translation feedback), but there were only 2 seemingly relevant out of 20 retrieved documents for the base query (used for pre-translation feedback). In the former case, El Norte was used while SJMN was used in the latter. SJMN is an American paper from an earlier time period. We might get better results using a collection that gives greater coverage.

Table 9. Best post-translation feedback results.

	ES-BASE				SE-BASE			
Fdbk Terms	0	5	5	5	0	20	20	30
Train. Docs	0	10	30	50	0	10	50	10
Average Precision:								
	0.0922	0.1252	0.1346	0.1359	0.0823	0.0910	0.0916	0.0913
% Change:		35.8	46.1	47.5	10.6	11.3	10.9	
Precision:								
5 docs:	0.2100	0.2600	0.2400	0.2400	0.2000	0.2500	0.1800	0.2300
10 docs:	0.2050	0.2300	0.2300	0.2350	0.2100	0.1950	0.1850	0.1800
15 docs:	0.2000	0.1967	0.2267	0.2300	0.1867	0.1900	0.1800	0.1800
20 docs:	0.1900	0.1875	0.2125	0.2200	0.1975	0.1975	0.1575	0.1800
30 docs:	0.1717	0.1750	0.1950	0.2000	0.1900	0.1633	0.1483	0.1617

3.4 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 combined method was most effective on the SE-BASE queries yielding up to a 51% improvement in average precision as shown in Table 10. The queries sets for other combined-feedback runs show similar results. As would be expected, both precision and recall are targeted by the combined method. The improvements occur because better query terms are added after the final feedback. Those terms tend to fine tune the query and de-emphasize inappropriate definitions. ES-BASE queries showed more than 40% improvement after combined feedback. Results are shown in Table 10.

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

Table 10. Best combined pre-translation and post-translation feedback results

	SE-BASE					ES-BASE				
Fdbk Terms	0	10	30	10	20	0	5	20	20	5
Train. Docs	0	50	10	30	50	0	30	20	30	20
Average Precision:										
	0.0823	0.1102	0.1150	0.1166	0.1242	0.0922	0.1372	0.1329	0.1375	0.1366
% Change:		34.0	39.7	41.7	51.0		48.8	44.2	49.2	48.2
Precision:										
5 docs:	0.2000	0.2300	0.2300	0.2400	0.2600	0.2100	0.2600	0.2700	0.2700	0.2400
10 docs:	0.2100	0.2100	0.2350	0.2200	0.2200	0.2050	0.2450	0.2450	0.2500	0.2500
15 docs:	0.1867	0.2033	0.1967	0.2200	0.2000	0.2000	0.2433	0.2367	0.2467	0.2400
20 docs:	0.1975	0.1875	0.1825	0.2075	0.2125	0.1900	0.2400	0.2200	0.2350	0.2375

4 Related Work

Salton [10] discussed cross-lingual retrieval as early as 1970, manually assigning thesaurus classes to the terms contained in a small collection of French documents and their translations. The thesaurus classes acted as an interlingua between the two sets of documents. Preliminary studies seemed promising, but CLIR was less effective for French queries and English documents: The variability of French caused the translations of English terms from a single class to be mapped to French terms spanning several other classes. Some terms had no correct mapping from French to English, and vice-versa. The test collection was also very small by current standards. No system was ever implemented, so it is unclear how such a system would perform in practice.

Landauer and Littman [9] proposed a method for cross-lingual retrieval. Latent Semantic Indexing (LSI) [6] was used to create a multidimensional indexing space for a small parallel corpus of English documents and their French translations. Their method has been successful at retrieving a query's translation. However, no reports of effectiveness on the traditional retrieval task have been reported. In more heterogeneous collections, language variability might require a much larger indexing space. Thus, the computational complexity of LSI brings into question the kind of performance one could expect on a more realistic multilingual collection. The method also relies on the use of parallel corpora.

Another method that relies on parallel and aligned corpora has been suggested by Dunning and Davis [5]. 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, by reduction of the document space, generates a translation matrix. Tests of the effectiveness of the method have been limited by its computational complexity.

More recently, Davis and Dunning[3, 4] have developed several other approaches to query translation for cross-lingual retrieval, all relying on a parallel corpus. In two methods, "translations" are performed by replacing English (Spanish) query terms with high frequency or statistically significant terms from the Spanish (English) side of the parallel corpus. A third uses Evolutionary Programming (EP) to optimize queries generated by one of the first two methods or via word-by-word MRD translation. The EP approach was the most effective, but results were disappointing, with each of the methods performing well below the baseline (word-by-word translation).

5 Conclusions

Ambiguity introduced via MRD translation leads to poor retrieval effectiveness. One of the two factors affecting this is the transfer of too many senses that are inappropriate to the query. The second factor is phrase loss caused by word-by-word translation of concepts that are more appropriately translated as phrases.

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 targets precision. Base queries which are less well specified (SE-BASE) show the greatest improvement. Local feedback after MRD translation targets recall. Improvements in performance are due to the introduction of terms which de-emphasize irrelevant translations and thus reduce ambiguity. Combining the two previous methods is also effective; short queries gain the greatest improvements.

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 automatically and do not rely on scarce resources such as parallel corpora. More importantly, they dramatically reduce translation error. Results show that half of the loss in performance can be regained with these methods. This appears to be mostly due to reducing the negative affects of inappropriate definitions. This approach is significantly better than similar approaches taken previously with realistic document collections.

Ambiguity arising from the word-by-word translation of phrases is one of the remaining factors 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. INFINDER [8], which automatically builds a corpus based association thesaurus, may also be useful for identifying the context in which certain words are used.

6 Acknowledgments

Access to the Collins bi-lingual MRDs was greatly facilitated by the use of tools such as NORM at the CRL lab at New Mexico State University. We would also like to thank Patty and Warren Greiff for their time in performing the manual translations. 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

1. R. Attar and A. S. Fraenkel. Local feedback in full-text retrieval systems. *Journal of the Association for Computing Machinery*, 24:397–417, 1977.
2. J. Broglio, J. Callan, , and W.B. Croft. Inquery system overview. In *Proceedings of the TIPSTER Text Program (Phase I)*, pages 47–67, 1994.
3. Mark Davis and Ted Dunning. Query translation using evolutionary programming for multi-lingual information retrieval. In *Proceedings of the Fourth Annual Conference on Evolutionary Programming*, 1995.
4. 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.
5. Ted Dunning and Mark Davis. Multi-lingual information retrieval. Technical report MCCS-93-252, Computing Research Laboratory, New Mexico State University, 1993.
6. 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.
7. Donna Harman, editor. *Proceedings of the 4th Text Retrieval Conference (TREC-4)*. 1996.
8. Y. Jing and W.B. Croft. An association thesaurus for information retrieval. In *RIAO 94 Conference Proceedings*, pages 146–160, 1994.
9. Thomas K. Landauer and Michael L. Littman. Fully automatic cross-language document retrieval. In *Proceedings of the Sixth Conference on Electronic Text Research*, 1990.
10. Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297, 1990.
11. Howard R. Turtle and W. Bruce Croft. Efficient probabilistic inference for text retrieval. In *RIAO 3 Conference Proceedings*, pages 664–661, 1991.
12. 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.

This article was processed using the LaTeX macro package with LLNCS style