Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources

Paul McNamee and James Mayfield
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, MD 20723-6099 USA
+1-443-778-3816
{mcnamee, mayfield}@jhuapl.edu

ABSTRACT

The quality of translation resources is arguably the most important factor affecting the performance of a cross-language information retrieval system. While many investigations have explored the use of query expansion techniques to combat errors induced by translation, no study has yet examined the effectiveness of these techniques across resources of varying quality. This paper presents results using parallel corpora and bilingual wordlists that have been deliberately degraded prior to query translation. Across different languages, translingual resources, and degrees of resource degradation, pre-translation query expansion is tremendously effective. In several instances, pre-translation expansion results in better performance when no translations are available, than when an uncompromised resource is used without pre-translation expansion. We also demonstrate that post-translation expansion using relevance feedback can confer modest performance gains. Measuring the efficacy of these techniques with resources of different quality suggests an explanation for the conflicting reports that have appeared in the literature.

1. INTRODUCTION

Cross-Language Information Retrieval (CLIR) systems seek to identify pertinent information in a collection of documents containing material in languages other than the one in which the user articulated her query. Intrinsic to the problem is a need to transform the query, document, or both, into a common terminological representation, using available translation resources. Thus, system performance is necessarily limited by the caliber of translations; clearly resources with broader coverage are preferable. High quality linguistic resources are typically difficult to obtain and exploit, or expensive to purchase. Participants in the major international CLIR evaluations such as CLEF, NTCIR, and TREC ([29], [30], [31]) frequently express a desire for better, and preferably low-cost, translation resources. The large multilingual collections available on the Internet have motivated researchers to attempt mining unstructured sources of linguistic data (e.g., Resnik [24]), fueled by the natural expectation that the use of more comprehensive resources will yield improvements in cross-language performance. It has even been suggested that CLIR evaluations may be measuring resource quality foremost (or equivalently, financial status) [7]. Scanning the papers of CLIR Track participants in TREC-9 and TREC-2001, we observe a trend toward the fusion of multiple resources in an attempt to improve lexical coverage. Clearly a need for enhanced resources is felt.

Typically, three types of resources are exploited for translingual mappings: bilingual wordlists (or machine readable dictionaries); parallel texts; and machine translation systems. The favorite appears to be bilingual wordlists, which are widely available, can be easy to use (especially if only word-by-word translation is attempted), and which preserve information such as alternate translations. Techniques using aligned parallel texts to produce statistical translation equivalents have become widely used since the publication of a method using Latent Semantic Indexing by Landauer and Littman [16]; however, these corpora are difficult to obtain and must first be aligned and indexed. Machine translation (MT) systems are perhaps the easiest approach for query translation, but may be computationally prohibitive for document translation. MT systems typically produce only a single candidate translation; thus some information of potential use to a retrieval system is lost. For an overview of translation methods in CLIR, see Oard and Diekema [19].

Regardless of the type of resource(s) used, several problems remain. Pirkola et al. [21] outline the major issues from a dictionary-based perspective; however, many of these same concerns arise when corpora or MT systems are used. They list difficulties with untranslatable terms, variations in inflectional forms, problems with phrase identification and translation, and translation ambiguity between the source and target languages as the main problems.

To cope with the paucity of translation resources and their inherent limitations, various techniques have been proposed. Query expansion is routinely used in monolingual retrieval, either by global methods such as thesauri, by local methods such as pseudo relevance feedback (PRF), or by local context analysis (LCA) [26]. In a multilingual setting, expansion can take place prior to translation, afterwards, or at both times.

The effect of resource quality on retrieval efficacy has received little attention in the literature. This study explores the relationship between the quality of a translation resource and CLIR performance. The effectiveness of both corpus and

SIGIR'02, August 11-15, 2002, Tampere, Finland.

dictionary-based resources was artificially lowered by randomly translating different proportions of query terms, simulating variability in the coverage of resources. We first discuss prior related work and then present our experimental design which explores multiple query expansion techniques. The remainder of the paper is devoted to an analysis of the empirical results.

2. PREVIOUS WORK

Regarding translation resources for CLIR, we believe that two points are widely agreed upon:

- resources are scarce and difficult to use; and
- resources with greater lexical coverage are preferable.

Because of the first point, the rarity of electronic sources for translation, investigators may be drawn to use the resources most readily available to them, rather than those best suited for bilingual retrieval. The second point is widely held, but to our knowledge, in only two cases has the benefit of increased lexical coverage been quantified [8], [27]; however, many different resources have been pair-wise compared extrinsically based on performance in bilingual retrieval tasks (e.g., [14], [18], [28]).

Degradation of documents and queries has been examined in two of the TREC evaluations, but only in a monolingual setting. Retrieval of garbled text documents was investigated to simulate a task where documents might contain numerous errors, such as if documents were created by optical character recognition [12]. And in TREC-9, short query forms containing realistic spelling errors were provided to test the ability of systems to cope with such mistakes. Also in TREC-9, the Query track examined the effects of query variability on system performance, but queries were re-stated, rather than purposefully weakened [5].

Query expansion based upon an entire query rather than on a candidate term's similarity to individual query search terms has been shown to be effective in monolingual settings [23]. Similarly, blind relevance feedback has been shown to be remarkably effective, especially when an initial query formulation lacks terms present in many relevant documents [25]. This might be the case when a query is very short, or when specific domain terminology (e.g., medicine, engineering) is used.

In a multilingual setting it seems plausible that pre-translation expansion would indeed be helpful. If a resource contains a restricted number of translatable search terms, then the degradation arising out of the translation process will cause many important query words to be unavailable for document ranking. But, if many words (or word forms) related to the query are translated, then the ultimate number of terms available for searching the target language is greater. This method presumes that the set of translated terms still represents the query semantics (i.e., the user's information request is not significantly altered by expansion *and* translation). If query translation does not produce a query with many coordinate terms, additional expansion through relevance feedback can likely improve precision as well as recall.

Many positive reports regarding the benefits of query expansion for CLIR have been reported; however, negative reports have been made frequently as well. We believe that differences in test collections, retrieval systems, language pairs, and translation resources obfuscate the conclusions of prior studies.

Ballesteros and Croft explored query expansion methods for CLIR and reported "combining pre- and post-translation expansion is most effective and improves precision and recall." [1] The use of both techniques led to an improvement from 42% to 68% of monolingual performance in mean average precision. The improvement from application of both methods was appreciably greater than the use of only pre- or post-translation expansion. Their work only examined a single language pair (English to Spanish), and relied on the Collins's English-Spanish electronic dictionary.

In a subsequent study [2], Ballesteros and Croft examined the use of co-occurrence statistics in parallel corpora to select translations from a machine-readable dictionary. Application of this technique was very effective and boosted bilingual performance from 68% to 88% of a monolingual baseline. Here they suggested that post-translation expansion helps remove errors due to incorrect translations.

More recently, Gey and Chen wrote an overview of the TREC-9 CLIR track, which focused on using English queries to search a Chinese news collection [9]. Their summaries of work by several top-scoring track participants reveal a disconcerting lack of consistency as to the merits of query expansion methods:

- 10% improvement in average precision with either pre-translation or post-translation expansion, but only short queries benefited from the use of both
- "Pre-translation query expansion did not help"
- "The best cross-language run did not use post-translation expansion"
- "Pre-translation expansion yielded an improvement of 42% over an unexpanded base run"
- "The best run used both pre- and post-translation expansion"
- "Post-translation query expansion yielded little improvement"

With inconsistent results like these, it is impossible to ascertain what techniques do and do not work. Each of the six systems referred to above used different translation resources, and we believe this amplifies the confusion. Until the effects of poor lexical coverage are better understood, shadows may hang over many research results unless the quality of translation resources employed is first ascertained. In an analysis of language resources used in the CLEF 2000 campaign [10], Gonzolo suggested measurement of resources and retrieval strategies in isolation, a recommendation we endorse.

In the TREC-2001 cross-language evaluation, which focused on English to Arabic retrieval, the system with the highest bilingual performance made use of several unique translation resources, which seems to agree with the notion that greater lexical coverage is helpful. However, it is impossible to discriminate between the benefits of the retrieval system that was employed and the resources utilized. Interestingly, the authors reported that pre-translation expansion was detrimental when post-translation relevance feedback was also applied, contradicting the results reported by Ballesteros and Croft [28].

A few investigations have examined the effect of resource size on CLIR performance. Two reports have measured retrieval performance as a function of resources for English-Chinese retrieval. Xu and Weischedel plotted performance on the TREC-5,6 Chinese tasks using a lexicon mined from parallel texts [27]. They used lexicons of a fixed size, where a lexicon of size *n* contained mappings for the *n* most frequent English words; bilingual performance was not improved for sizes greater than 20,000 terms. Franz et al. examined three parallel collections for use on the TREC-9 Chinese topics [8]. Using short queries, they found that out-of-vocabulary rate was more important than domain, dialect, or style in predicting system performance.

For the CLEF-2001 workshop, Kraaij examined the relative merits of an MT system, a lexical database, and a parallel corpus, and emphasized the benefits that can be obtained from combining such disparate translation resources [14]. With the use of all three resources he observed bilingual performance 98% of a monolingual baseline for English to French retrieval. Separate use of a dictionary, a corpus, and an MT system yielded performance only 73%, 90%, and 92% of a monolingual baseline. He offered the opinion that "the mean average precision of a run is proportional to the lexical coverage [of the translation resources]", but this statement appears to be based only on a qualitative examination of why performance on certain topics differed depending on the resources and language pairs used.

The results reported in the present paper confirm Kraaij's conjecture and quantify the degree to which inferior resource quality affects CLIR performance and under which circumstances query expansion techniques can mitigate translation errors due to poor lexical coverage.

3. EXPERIMENTS

3.1 Test Collection

The CLEF-2001 test collection was used for all of our experiments (see [20] for a description). The collection contains roughly 1 million newspaper articles published in 1994 or 1995 (see Table 1).

Table 1. CLEF-2001 Document Collection

	Documents	Unique words
Dutch	190,604	692,745
English	110,282	235,710
French	87,191	479,682
German	225,371	1,670,316
Italian	108,578	1,323,283
Spanish	215,737	382,664

The Bilingual Track in the CLEF-2001 evaluation permitted a variety of query languages to be used to search either the Dutch or English collections. Here we only explored the five language pairs Dutch, French, German, Italian, and Spanish, to English. The test suite contains fifty topic statements, but only forty-seven of the topics contain a relevant English article. A mixture of topics including local, national, and international subjects was selected. In each language topic statements were crafted by native speakers and significant effort was expended to ensure that the intended topic semantics were preserved in the respective languages.

3.2 Document and Query Processing

Document processing was designed to require minimal use of language specific resources such as stopword lists, lexicons, decompounders, stemmers, lists of phrases, or manually-built thesauri, so each language's sub-collection was handled much the same. Punctuation was eliminated, letters were down-cased, and diacritical marks were preserved. Thus documents and queries are represented as bags of unnormalized word forms. Queries were tokenized in the same fashion as documents, but obvious query structure (e.g., 'find documents that' or 'relevant documents must contain') was removed. We used a retrieval system developed in-house for all of our experiments. The system uses a statistical language model of retrieval with Jelinek-Mercer smoothing of document term frequencies. See [3], [13], and [22] for more details on these models.

To perform pre-translation expansion, we relied solely on local methods based on an initial retrieval from the appropriate source language sub-collection of the CLEF-2001 documents. For example, to investigate pre-translation expansion for Italian to English retrieval, we would first do a monolingual retrieval in the Italian collection (i.e., La Stampa and SDA-IT). Using the top ranked 25 retrieved documents as positive exemplars and presuming the lowest 75 ranked out of 1000 were irrelevant, we produced a set of 60 weighted terms for each query that included the original query terms; this is analogous to both query expansion and query term re-weighting as described in Harman [11]. It should be pointed out that the sub-collections in each language of the CLEF-2001 evaluation are contemporaneous, so this set of expansion terms might be somewhat better than an arbitrary monolingual collection. We did not investigate global methods for query expansion in the source language because this would have required a thesaurus in each source language that we wished to investigate.

When a query was expanded after translation, we again relied on pseudo relevance feedback based on terms extracted from retrieved target language documents. As with pre-translation expansion, we identified 60 weighted terms for use as an expanded query and searched the target language (English) collection for a second time.

3.3 Translation Resources

For reasons of convenience we only examined corpus- and dictionary-based translation – it was not clear to us how to best degrade commercial translation software since many packages are optimized for grammatically correct sentences rather than word-by-word translation. Both the parallel corpus and the multilingual wordlist were extracted from the Web. These resources were not validated and may contain numerous errors.

We collected a variety of bilingual wordlists where English was one of the languages involved. Translation equivalents for over ninety thousand English words are available in at least one of forty or so languages. We did not attempt to utilize or reverse engineer web-based interfaces to dictionaries, but rather only sought wordlists in the public domain, or whose use appeared unrestricted; the Ergane dictionaries [32] and files from the Internet Dictionary Project [34] are the largest sources. We used the ABET extraction tool to convert these disparate wordlists to machine-readable form [17].

When translating a word using a bilingual wordlist we simply use all of the alternative mappings for the word, and each mapping is weighted using the same query term frequency as the original word. In our wordlist the mean number of entries per term by language is: 3.01 for Dutch; 2.08 for French; 1.58 for German; 1.52 for Italian; and 1.57 for Spanish.

We also built a set of aligned corpora using text mined from the Europa site [33]; specifically, we downloaded eight months of the Official Journal of the European Union (December 2000 through August 2001). The Journal is published in eleven languages in PDF format. We converted the PDF formatted documents to text encoded in ISO-8859-1, aligned the documents using simple rules for whitespace and punctuation with Church's char_align program [6], and then indexed the data. It was easiest to construct a separate aligned corpus for each non-English language, rather than to build a single, multiply aligned collection. The resulting collection contains roughly 100MB of text in each language. The number of words with at least one English translation produced by the two resources is shown in Table 2. It should be noted that many of the terms extracted from the aligned corpus are names or numbers that would not normally be contained in a dictionary, so the number of entries reported here is not a clear indication of a superior resource.

Table 2. Bilingual Resource Size (in terms)

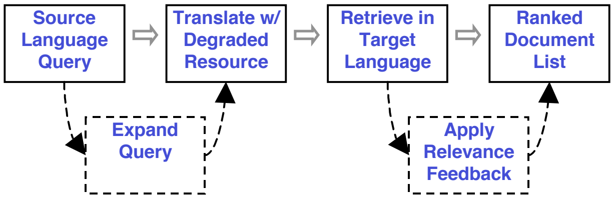
	Wordlist	Corpus
Dutch	15,591	184,506
French	23,322	135,454
German	94,901	224,961
Italian	18,461	138,890
Spanish	25,028	146,938

When translating a word using an aligned corpus, we select the single best candidate translation.

3.4 Experimental Design

We now describe the experiments we undertook. We focused only on word-by-word query translation because of its simplicity. Our goal is to compare four methods of query expansion or augmentation under a spectrum of conditions corresponding to differing quality translation resources. The four methods examined are no use of expansion, pre-translation expansion only, post-translation only, and the use of both pre- and post-translation expansion. Figure 1 illustrates the procedure we followed.

Previously we mentioned that only 47 of the CLEF-2001 topics contain a relevant English article; however, 12 additional topics contain only one or two relevant documents. This may be attributable to the design goals of the evaluation: a certain number topics were sought that focused on local subjects, and the American-based LA Times is less likely to report on these issues. Since relevance feedback is only expected to enhance retrieval performance when a reasonable number of germane documents are present in the target language collection, we chose to evaluate our runs using the 35 topics with three or more relevant documents. Topics 44, 52, 54, 57, 59, 60, 62, 63, 67, 73, 74, 75, 78, 79, and 88 were discarded. Kwok and Chan [15] developed a technique designed to provide for good query expansion in this situation (where a target collection only has a small number of relevant documents), but we did not attempt it here. Their idea is based on searching a larger collection that is expected to contain many more documents about that domain; they termed the technique 'collection enrichment'.



Title, Title + Desc, and Title + Desc + Narr query forms were considered in each of five source languages: Dutch, French, German, Italian, and Spanish. Source language expansion or no expansion was used, followed by translation. There were two methods for translation: use of a bilingual dictionary using all available translations; and statistical translation using an aligned parallel corpus. Eleven versions of each resource were simulated corresponding to different measures of lexical coverage. After translation, retrieval was performed on the target language collection, which was English, and optionally, pseudo relevance feedback was applied. A total of 1320 runs were created.

Figure 1. Overview of Experiments Performed

We considered two methods for impairing our translation resources. The first method was the simple idea of physically creating new wordlists or corpora with missing lexical entries. This seemed laborious, so instead we opted for simulating weaker resources by randomly declining to translate a given percentage of query terms. In other words, for each term, we would generate a random number between 0 and 1, and only if the value was greater than the degree of degradation did we attempt to find a target language mapping. We could have removed a percentage of all lexical entries from the resource, but since only a small percentage of the terms occur in the CLEF queries, this would be counterproductive. The same random seeds were used for both corpus or wordlist translations. In practice a language resource would likely have more mappings for common terms and fewer entries for proper nouns or obscure terms. We did not attempt to model this, but dropping low frequency words is probably a better idea than randomly omitting query terms.

Starting with no degradation, we removed terms in increments of 10%, up to complete degradation. When a decision was made not to translate a given term, the untranslated form was left in the query as a potential translation. This is a common practice, and is motivated by the observation that in related languages, many morphological cognates exist. Thus, even when a resource is 100% degraded, corresponding to a state in which no translation resource is available, it is still possible to retrieve relevant documents.

Different terms will be omitted from a query for a particular random seed; this is expected to increase the variance in our evaluation measures. Averaging over a number of trials, each using a different seed, would provide a clear solution to this problem. We decided against this for reasons of expediency; otherwise the number of runs would have been unmanageable. We chose to focus primarily on mean average precision to evaluate our results, but we collected statistics for precision at low recall levels as well.

In Table 3 a monolingual baseline is compared to bilingual queries at four levels of resource impairment and the effect of pre-translation expansion is shown. Italian is used as the source language and the parallel corpus is used to map terms into English for the short version of query 66, "Russian Withdrawal from Latvia" (Italian: "Ritiro delle truppe russe dalla Lettonia "). At 0% degradation pre-translation expansion is dramatically better due to several poor translations; at 40%, these translations are dropped because of the random resource degradation, so performance actually rises; however, at 80% only the term 'withdrawal' is correctly translated without expansion, so expansion is critical here. We note that at 100% degradation we still obtain a reasonable degree of performance, but only when expansion is used. This is due to cognates (like "estonia" and "russia") that were extracted from Italian articles during source language query expansion, but which require no translation into English.

Table 3. Illustration of the effects of pre-translation expansion and resource degradation

Degradation / Query	Recall at 1000 docs	Average Precision	Precision at 10 docs
English Monolingual	11	0.6923	0.8
{latvia=1, russian=1, withdrawal=1}			
0% Degradation	11	0.1176	0.1
{communities=1, directive=1, latvia=1, russian=1, troops=1, withdrawal=1}			
With Expansion	11	0.4791	0.7
{1517#=80, 31=119, agency=80, agreement=83, agreements=103, armed=74, august=134, baltic=177, countries=112, estonia=144, foreign=95, incubators=73, latvia=76, latvian=135, line=215, lithuania=135, living=82, maintain=92, military=112, minorities=77, moscow=199, near=71, negotiations=78, news=109, north=76, pension=73, pensioners=71, press=85, radar=94, reported=76, rights=86, russia=212, russian=74, service=84, soldiers=108, station=84, suspended=76, tallinn=77, troops=926, unit=70, voltage=75, warsaw=101, withdrawal=958, within=135}			
40% Degradation	11	0.6982	0.8
{delle=1, latvia=1, russian=1, troops=1, withdrawal=1}			
With Expansion	11	0.4400	0.5
{31=119, agosto=134, agreement=83, baltic=177, estonia=144, incubators=73, latvia=76, latvian=135, line=215, maintain=92, military=112, moscow=199, near=71, negotiations=78, news=109, north=76, pension=73, radar=94, reported=76, rights=86, russia=212, russian=74, service=84, soldiers=108, stampa=85, station=84, suspended=76, tallinn=77, troops=926, unit=70, voltage=75, warsaw=101, withdrawal=958, within=135}			
80% Degradation	11	0.0069	0.0
{delle=1, directive=1, russe=1, withdrawal=1}			
With Expansion	11	0.5077	0.6
{31=119, agosto=134, baltic=177, dislocate=73, estone=144, estonia=238, near=71, negotiations=78, news=109, pensione=73, radar=94, reported=76, riga=215, russe=1019, russia=212, russian=74, russo=118, service=84, stampa=85, station=84, tallinn=77, troops=926, unit=70, withdrawal=958}			
100% Degradation	0	0.0000	0.0
{dalla=1, delle=1, russe=1}			
With Expansion	11	0.2401	0.2
{31=119, agosto=134, baltic=148, dislocate=73, estone=144, estonia=238, news=109, pensione=73, radar=94, riga=215, russa=74, russe=1019, russi=193, russia=212, russo=118, service=84, stampa=85, stazione=84, tallinn=77}			

4. RESULTS

The resources used for translation in our experiments are uncurated resources derived from the Web. Because the adequacy of these resources for cross-language retrieval has not previously been demonstrated, we first assessed the performance of the uncompromised resources. Only if a sufficient level of performance was seen would our experiments be meaningful; otherwise concern about whether these conclusions hold for superior resources would arise.

A baseline of English monolingual performance is shown in Table 4, for the three query forms (title-only (or T), title+description (or TD), and title+description+narrative (or TDN)) with and without the application of pseudo relevance feedback.

Table 4. Mean Average Precision of a Monolingual Baseline

	T	T w/RF	TD	TD w/RF	TDN	TDN w/RF
English	0.3578	0.4067	0.4383	0.4284	0.4825	0.4780

In Table 5 we report the percentage of mean average precision achieved by each bilingual run performed with intact translation resources when pre-translation expansion was not used. For our English baselines, relevance feedback improved the title-only queries, but did not appreciably change when longer topic statements were used. Each column in the table (below) is compared to the corresponding English run. We observe that when the parallel corpus is used for translation, an average of between 68% and 75% relative performance is obtained, depending on the run condition; with our dictionary, only 35% to 59% is seen on average. The dictionary appears to be an inferior resource, but the lower performance could also be attributable to our failure to normalize word forms. Longer topic statements fare better, and relevance feedback is somewhat helpful. We point out that pre-translation query expansion was not used in the table below. Given the lower performance when using the dictionary for translation, we must be cautious in drawing conclusions from those data.

Table 5. Bilingual Performance with Uncompromised Resources (percentage of monolingual performance)

		T	T w/RF	TD	TD w/RF	TDN	TDN w/RF
Dutch	Corp.	66.6	67.4	63.0	71.6	59.7	61.5
	Dict.	43.9	55.2	26.3	35.3	24.5	41.4
French	Corp.	59.9	67.5	73.0	80.9	73.8	86.9
	Dict.	57.2	57.4	48.7	60.7	61.3	78.9
German	Corp.	70.9	69.6	65.4	72.3	67.0	69.0
	Dict.	42.3	37.5	26.1	33.3	38.7	43.0
Italian	Corp.	70.5	70.6	72.5	80.5	75.4	78.2
	Dict.	35.5	51.0	33.4	48.2	39.6	61.2
Spanish	Corp.	72.4	69.7	72.2	75.9	75.6	81.4
	Dict.	51.2	51.8	41.2	54.2	57.6	71.6
Mean	Corp.	68.1	69.0	69.2	72.2	70.3	75.4
	Dict.	46.0	50.6	35.1	46.3	44.3	59.2

Now we get to the heart of the matter – addressing the question of how performance worsens as a translation resource is degraded. Figure 2 shows the performance in an agglutinative language, Dutch; and retrieval in Spanish is illustrated in Figure 3. For each language, six conditions are shown corresponding to the use of T, TD, or TDN topic statements with either corpus- or dictionary-based translation.

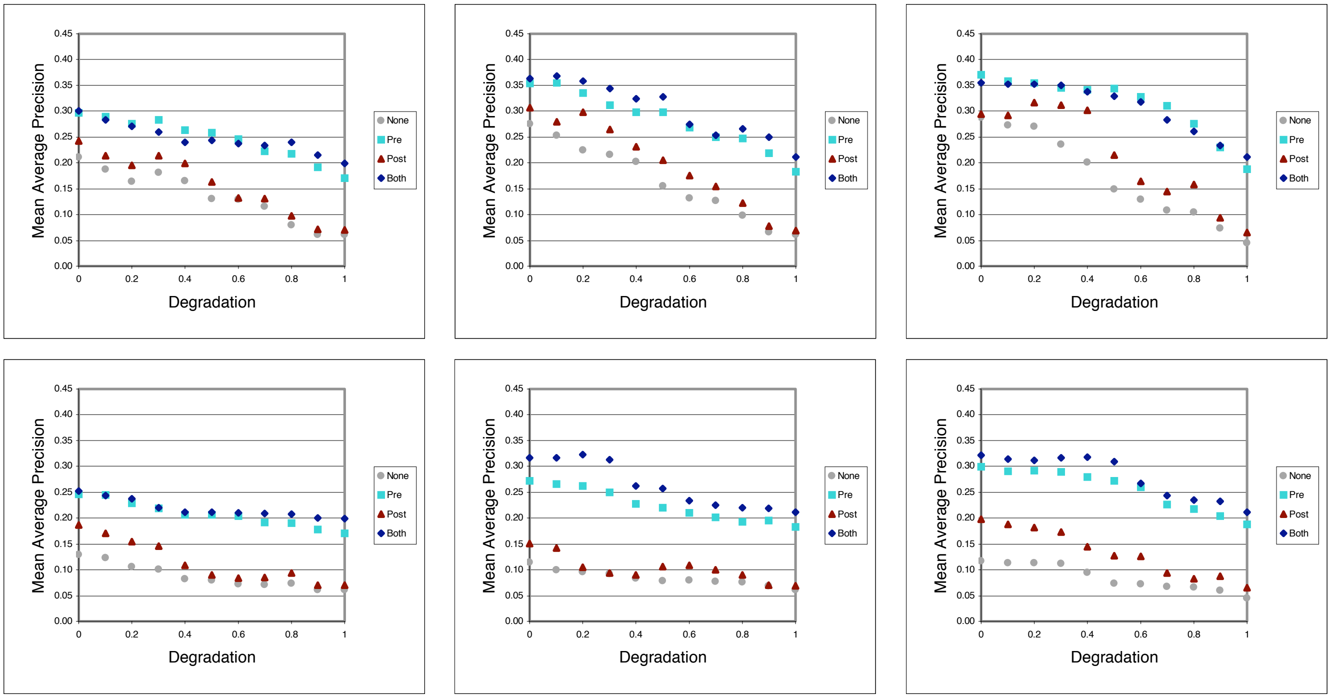


Figure 2. Effectiveness of expansion techniques as a function of resource degradation for the Dutch topics. Going from left to right, the three plots on the top row used the title-only, title+description, and title+description+narrative topic statements, respectively, and the parallel corpus for translation. Dictionary-based translation was used for the plots on the second row. Each plot shows the performance under four conditions: no expansion; only pre-translation expansion; only post-translation expansion; and both pre- and post-translation expansion.

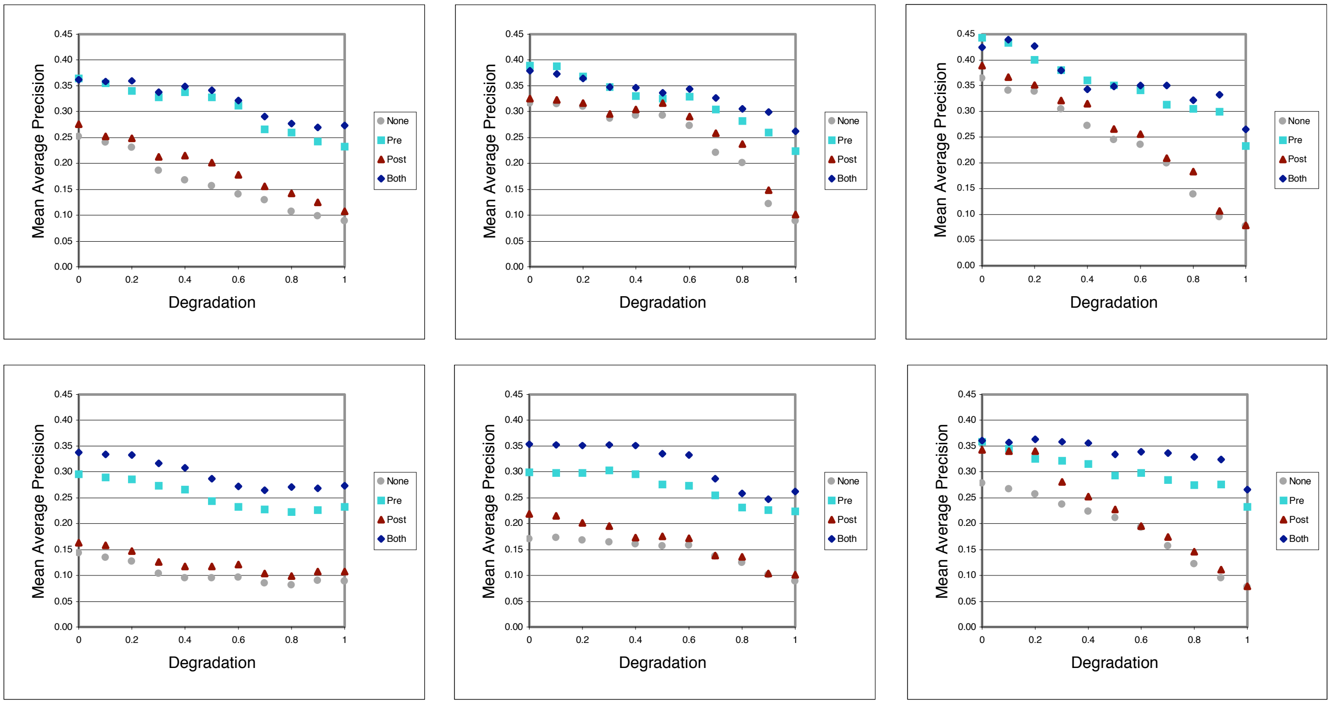


Figure 3. Effectiveness of expansion techniques as a function of resource degradation for the Spanish topics. The plots are arranged as in the previous figure.

4.1 No Expansion

Looking at Figures 2 and 3, we first note that retrieval performance drops linearly with decreased lexical coverage when no expansion is performed, confirming Kraaij's conjecture. The decrease depends on the caliber of the resource (the dictionary plots are noticeably worse), and on the length of the query. Unsurprisingly, longer queries perform better: they have further to fall when a weaker resource is used.

4.2 Post-Translation Expansion Alone

We find that the use of blind relevance feedback consistently increases the mean average precision by a modest amount. This occurs in each of the five language pairs and across variations in the lexical coverage of the different translation resources.

4.3 Pre-Translation Expansion Alone

Pre-translation expansion is tremendously useful across all levels of degradation. At higher levels of degradation, gains between 200 and 300% are realized. Only when a comprehensive translation resource is used, or when no comparable expansion collection is available, would we expect to see no benefit from expansion. Therefore, we recommend that this technique be applied whenever gains in precision justify the computational and procedural complexity of automated query expansion.

Amazingly, with no resource at all (i.e., the situation when a resource is 100% degraded), pre-translation expansion alone can result in better performance than when an uncompromised resource is used without expansion. This follows earlier work by Buckley et al. [4] that viewed English as "misspelled French" and attempted bilingual retrieval using rules for spelling correction and reliance on cognate matches. Pre-translation expansion appears to multiply the number of cognates useful for retrieval in related languages.

4.4 Pre- and Post-Translation Expansion

Finally, in agreement with the work cited earlier by Ballesteros and Croft, we confirm that a combination of pre- and post-translation expansion often yields the greatest performance. However, pre-translation expansion is responsible for the greatest gains. We see an improvement of approximately 10% - 15% when relevance feedback is also applied. This occurs either when the inferior resource, the wordlist, is used, or at high levels of degradation when the parallel corpus is used for translation.

4.5 Results in Other Languages

Figures 2 and 3 illustrated the detriment that occurs when a weaker translation resource is used, along with the ability of query expansion to ameliorate the losses due to poor lexical coverage, in Dutch and Spanish. The same trends hold in French, German, and Italian. A comparison of expansion techniques at four levels of lexical coverage is shown in Table 6.

The table shows the mean average precision experienced with corpus-based translation and TD topics. The highlighted cells indicate when an increase in performance using an expansion technique was statistically significant at the 95% confidence level (Wilcoxon test). The use of both pre-translation and post-translation expansion is almost always better, but at low levels of degradation, pre-translation expansion alone sometimes outperforms the combination. With high quality resources, many of the expansion terms will be correctly translated, and so gains that normally would occur by finding words related to, but not present in the initial query, using relevance feedback, are found instead by the initial feedback from the source language.

Table 6. Effects of Corpus Degradation on Expansion Utility

		0%	30%	70%	100%
Dutch	None	0.2759	0.2168	0.1274	0.0623
	Pre	0.3537	0.3121	0.2502	0.1832
	Post	0.3067	0.2643	0.1548	0.0697
	Both	0.3640	0.3439	0.2529	0.2113
French	None	0.3199	0.2502	0.2028	0.1091
	Pre	0.3603	0.3264	0.2618	0.1927
	Post	0.3467	0.2964	0.2907	0.1451
	Both	0.3698	0.3368	0.3060	0.2362
German	None	0.2784	0.2352	0.1409	0.0727
	Pre	0.3290	0.2903	0.2628	0.1944
	Post	0.3009	0.2566	0.1717	0.1135
	Both	0.3448	0.2974	0.3043	0.2440
Italian	None	0.3178	0.2482	0.2061	0.0823
	Pre	0.4209	0.3754	0.2552	0.2012
	Post	0.3449	0.3005	0.2310	0.0899
	Both	0.3989	0.3528	0.2643	0.2292
Spanish	None	0.3164	0.2863	0.2219	0.0887
	Pre	0.3890	0.3478	0.3040	0.2233
	Post	0.3253	0.2950	0.2583	0.1018
	Both	0.3799	0.3479	0.3262	0.2620

4.6 Limitations

To consider a breadth of source languages, query lengths, and expansion methods, some compromises were made; these should be considered in evaluating our results. Such factors include using a simple method for query term translation (unbalanced translation without translation of multiword units), reliance on contemporaneous newsprint collections for expansion, and use of a single random seed when selecting query terms not to translate.

5. CONCLUSIONS

In this paper we have demonstrated empirically the intuitive notion that bilingual retrieval performance drops off as the lexical coverage of translation resources decreases, and we confirmed that the relationship is approximately linear. Moreover, by using degraded translation resources we presented a framework to discover under which circumstances traditional query expansion techniques prove most beneficial.

We strongly recommend the use of pre-translation expansion when dictionary- or corpus-based query translation is performed; in some instances this expansion can treble performance. However, the computational expense and availability of comparable expansion collections should be considered. Additional relevance feedback in the target language is often useful, and can provide an additional 10-15% benefit. However, when high quality (i.e., comprehensive) resources are available, little gain is likely to occur. Differences in resource quality may account for disagreeing reports on the effectiveness of query expansion in cross-language retrieval.

We also demonstrated that even with very poor cross-language resources, good performance is still feasible when pre-translation expansion is used. This result is particularly important because it suggests that translingual retrieval in low-density languages will benefit significantly from such expansion.

6. REFERENCES

[1] L. Ballesteros and W. B. Croft, 'Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval.' In the *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-97),* pp. 84-91, 1997.

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

[3] A. Berger and J. Lafferty, 'Information Retrieval as Statistical Translation.' In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99)*, pp. 222-229, 1999.

[4] C. Buckley, M. Mitra, J. Walz, and C. Cardie, 'Using Clustering and Super Concepts within SMART: TREC-6.' In E. Voorhees and D. Harman (eds.), *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, NIST Special Publication 500-240, 1998.

[5] C. Buckley, 'The TREC-9 Query Track.' In E. M. Voorhees and D. K. Harman, eds., *Proceedings of the Ninth Text REtrieval Conference (TREC-9),* pp. 81-85, 2001.

[6] K. W. Church, 'Char_align: A program for aligning parallel texts at the character level.' In the *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 1-8, 1993.

[7] A. Diekema, '*May the Best Team Win: Language Resources in CLIR*.' Position paper at the CLEF-2000 workshop. Available online at: http://clef.iei.pi.cnr.it:2002/DELOS/CLEF/diekema.pdf

[8] M. Franz, J. S. McCarley, T. Ward, and W. Zhu, 'Quantifying the Utility of Parallel Corpora.' *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-01)*, pp. 398-399, 2001.

[9] F. Gey and A. Chen, 'TREC-9 Cross-Language Information Retrieval (English – Chinese) Overview.' In E. M. Voorhees and D. K. Harman, eds., *Proceedings of the Ninth Text REtrieval Conference (TREC-9),* pp. 15-23, 2001.

[10] J. Gonzolo, 'Language Resources in Cross-Language Text Retrieval: A CLEF Perspective.' In Carol Peters (ed.) *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF-2000 Workshop*, Lisbon, Portugal, Lecture Notes in Computer Science 2069, Springer, pp. 36-47, 2001.

[11] D. Harman, 'Relevance Feedback Revisited.' In the *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-92)*, pp. 1-10, 1992.

[12] D. Harman, 'Overview of the Fourth Text REtrieval Conference (TREC-4).' In D. K. Harman, ed., *Proceedings of the Fourth Text REtrieval Conference (TREC-4),* NIST Special Publication 500-236, pp. 1-24, 1995.

[13] D. Hiemstra, '*Using Language Models for Information Retrieval*.' Ph. D. Thesis, Center for Telematics and Information Technology, The Netherlands, 2000.

[14] W. Kraaij, 'TNO at CLEF-2001: Comparing Translation Resources.' To appear in Carol Peters (ed.), *Proceedings of the CLEF-2001 Workshop*.

[15] K. L. Kwok and M. Chan, 'Improving Two-Stage Ad-Hoc Retrieval for Short Queries.' In the *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98)*, pp. 250-256, 1998.

[16] T. K. Landauer and M. L. Littman, 'Fully automated cross-language document retrieval using latent semantic indexing.' In the *Proceedings of the 6th Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*. 31-38, 1990.

[17] J. Mayfield and P. McNamee, 'Converting On-Line Bilingual Dictionaries from Human-Readable to Machine-Readable Form.' In these proceedings.

[18] P. McNamee and J. Mayfield, 'JHU/APL Experiments at CLEF: Translation Resources and Score Normalization.' To appear in Carol Peters (ed.), *Proceedings of the CLEF-2001 Workshop*.

[19] D. Oard and A. Diekema, 'Cross-Language Information Retrieval.' In M. Williams (ed.), *Annual Review of Information Science*, pp. 223-256, 1998.

[20] C. Peters, 'Foreward to the *Proceedings of the CLEF-2001 Workshop*', to appear in 2002.

[21] A. Pirkola, T. Hedlund, H. Keskusalo, and K. Järvelin, 'Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings.' In *Information Retrieval*, vol. 4, pp. 209-230, 2001.

[22] J. M. Ponte and W. B. Croft, 'A Language Modeling Approach to Information Retrieval.' In the *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98)*, pp. 275-281, 1998.

[23] Y. Qiu and H. P. Frie, 'Concept Based Query Expansion.' In the *Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-93),* 1993.

[24] P. Resnik, 'Mining the Web for Bilingual Text.' *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 1999.

[25] G. Salton and C. Buckley, 'Improving Retrieval Performance by Relevance Feedback.' In the *Journal of the American Society for Information Science*, 41(4), pp. 288-297, 1990.

[26] J. Xu and W. B. Croft, 'Query Expansion Using Local and Global Document Analysis.' In the *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-96)*, pp. 4-11, 1996.

[27] J. Xu and R. Weischedel, 'Cross-lingual Information Retrieval Using Hidden Markov Models.' In the *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000),* 2000.

[28] J. Xu, A. Fraser, and R. Weischedel, 'TREC 2001 Cross-lingual Retrieval at BBN.' In *TREC-2001 Notebook Papers*, pp. 122-131, 2001.

[29] Cross-Language Evaluation Forum, http://www.clef-campaign.org/

[30] NTCIR Project, http://research.nii.ac.jp/ntcir/

[31] Text REtrieval Conference, http://trec.nist.gov/

[32] http://dictionaries.travlang.com/

[33] http://europa.eu.int/

[34] http://www.june29.com/IDP/