Scalable Multilingual Information Access

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

The third Cross-Language Evaluation Forum workshop (CLEF-2002) provides the unprecedented opportunity to evaluate retrieval in eight different languages using a uniform set of topics and assessment methodology. This year the Johns Hopkins University Applied Physics Laboratory participated in the monolingual, bilingual, and multilingual retrieval tasks. We contend that information access in a plethora of languages requires approaches that are inexpensive in developer and run-time costs. In this paper we describe a simplified approach that seems suitable for retrieval in many languages; we also show how good retrieval is possible over many languages, even when translation resources are scarce, or when query-time translation is infeasible. In particular, we investigate the use of character n-grams for monolingual retrieval, pre-translation expansion as a technique to mitigate errors due to limited translation resources, and translation of document representations to an interlingua for computationally efficient retrieval against multiple languages.

Introduction

The number of languages in the CLEF document collection has grown to eight in 2002: Dutch, English, Finnish, French, German, Italian, Spanish, and Swedish. While the Romance languages have a great deal in common with one another, the Teutonic languages and Finnish have different origins; this set of modern languages provide challenges in word decompounding, complex morphology, and handling diacritical marks. For many years research in information retrieval was focused on the English language where these problems are less significant. As a result simple rules for stemming words and case-folding are really the only common improvements to exact string matching used by retrieval systems. The use of stopword lists is also routine, but seems to have little effect except to reduce the size of inverted files and to improve runtime efficiency.

We have been interested in discovering how simple methods can be applied to combat the aforementioned problems. Though their use has not found favor in English, we have demonstrated that overlapping character n-grams are remarkably effective for retrieval in many languages, including those most widely used in Europe. This simple technique appears to provide a surrogate means to normalize word forms, an efficient approximation to word bigrams (when n-grams with interior spaces are formed), and a solution to the problem of decompounding agglutinative languages. For the CLEF-2002 evaluation we continued to use the Hopkins Automated Information Retriever for Combing Unstructured Text (HAIRCUT) system which supports n-gram processing.

We participated in three tasks at this year's workshop, monolingual, cross-language, and multilingual retrieval. All of our official submissions were automated runs. This year we relied on an aligned parallel corpus as our sole translation resource – this resource was automatically mined from the Web and can be used to support retrieval between any pair of E.U. languages, except Greek. In the sections that follow, we first describe the standard methodology used for each language's sub-collection and we then present initial results in monolingual, bilingual, and multilingual retrieval. Highlights include an investigation into the use of pre-translation expansion from a comparable collection to improve retrieval performance, a discovery that character n-grams provide a means for effective bilingual retrieval for a close language pair, without translation, and an efficient method for multilingual retrieval that involves no query-time translation.

Methodology

For the monolingual tasks we created sixteen indexes, a word and an n-gram (n=6) index for each of the eight languages. For the bilingual and multilingual tasks we used the same indexes but translated topic statements to produce our official runs; however, we also report on another approach for multilingual retrieval that required a separate index. Information about each index is provided in Table 1.

	# docs	collection size (MB zipped)	type	# terms	index size (MB)
Dutch	190,604	203	words	692,754	160
			6-grams	3,816,580	1133
English	110,282	166	words	235,713	98
			6-grams	2,944,813	889
Finnish	55,344	51	words	981,174	87
			6-grams	2,524,529	383
French	87,191	92	words	248,225	68
			6-grams	2,343,009	511
German	225,371	207	words	1,079,453	221
			6-grams	4,203,047	1,325
Italian	108,578	107	words	338,634	89
			6-grams	2,162,249	607
Spanish	215,737	186	words	382,666	150
			6-grams	3,193,404	1098
Swedish	142,819	94	words	510,245	95
			6-grams	3,254,595	628

Table 1. Information about indexes for the CLEF-2002 test collection

Index Construction

Our methods for scanning documents, creating an index, and processing queries are essentially unchanged from last year. We include below a description from our CLEF-2001 paper [3]; those already familiar with our our previous work using HAIRCUT should skip ahead to a description of this year's experiments.

Documents were processed using only the permitted tags specified in the workshop guidelines. First SGML macros were expanded to their appropriate Unicode character. Then punctuation was eliminated, letters were downcased, and only the first four of a sequence of digits were preserved (e.g., 010394 became 0103##). Diacritical marks were preserved. The result is a stream of words separated by spaces. Exceedingly long words were truncated; the limit was 35 characters in the Dutch and German languages and 20 otherwise. When using n-grams we extract indexing terms from the same stream of words; thus, the n-grams may span word boundaries, but sentence boundaries are noted so that n-grams spanning sentence boundaries are not recorded. N-grams with leading, central, or trailing spaces are formed at word boundaries. For example, given the phrase, "the prime minister," the following 6-grams are produced.

Term	Document Frequency	Collection Frequency	IDF	RIDF
-the-p	72,489	241,648	0.605	0.434
the-pr	41,729	86,923	1.402	0.527
he-pri	8,701	11,812	3.663	0.364
e-prim	2,827	3,441	5.286	0.261
-prime	3,685	5,635	4.903	0.576
prime-	3,515	5,452	4.971	0.597
rime-m	1,835	2,992	5.910	0.689
ime-mi	1,731	2,871	5.993	0.711
me-min	1,764	2,919	5.966	0.707
e-mini	3,797	5,975	4.860	0.615
-minis	4,243	8,863	4.699	1.005
minist	15,428	33,731	2.838	0.914
iniste	4,525	8,299	4.607	0.821
nister	4,686	8,577	4,557	0.816
ister-	7,727	12,860	3.835	0.651

Table 2. Example 6-grams produced for the input "the prime minister." Term statistics are based on the LA Times subset of the English collection. Dashes indicate whitespace characters.

The use of overlapping character n-grams provides a surrogate form of morphological normalization. For example, in Table 2 above, the n-gram "minist" could have been generated from several different forms like *administer*, *administrative*, *minister*, *ministers*, *ministerial*, or *ministry*. It could also come from an unrelated word like *feminist*. Another advantage of n-gram indexing comes from the fact that n-grams containing spaces can convey phrasal information. In the table above, 6-grams such as "rime-m", "ime-mi", and "me-min" may act much like the phrase "prime minister" in a word-based index using multiple word phrases.

At last year's workshop we examined different types of translation resources for bilingual retrieval and espoused a language-neutral approach to retrieval. We continued in this vein and did not utilize stopword lists or morphological analyzers.

Query Processing

HAIRCUT performs rudimentary preprocessing on topic statements to remove stop structure, *e.g.*, phrases such as "… would be relevant" or "relevant documents should…." . We have constructed a list of about 1000 such English phrases from previous topic sets (mainly TREC topics) and these have been translated into other languages using commercial machine translation. Other than this preprocessing, queries are parsed in the same fashion as documents in the collection.

In all of our experiments we used a linguistically motivated probabilistic model for retrieval. Our official runs all used blind relevance feedback, though it did not improve retrieval performance in every instance. To perform relevance feedback we first retrieved the top 1000 documents. We then used the top 20 documents for positive feedback and the bottom 75 documents for negative feedback; however, we removed any duplicate or near duplicate documents from these sets. We then select terms for the expanded query based on three factors, a term's initial query term frequency (if any); the cube root of the (α=3, β=2, γ=2) Rocchio score; and a term similarity metric that incorporates IDF weighting. The 60 top ranked terms are then used as the revised query with words as indexing terms; 400 terms are used with 6-grams. In previous work we penalized documents containing only a fraction of the query terms; we are no longer convinced that this technique adds much benefit and have discontinued its use. As a general trend we observe a decrease in precision at very low recall levels when blind relevance feedback is used, but both overall recall and mean average precision are improved.

Monolingual Experiments

We submitted an official run for each target language only using the <title> and <desc> fields and only automatic processing. These official runs were actually the combination of two base-runs, one using words and one using 6-grams; both base-runs also make use of blind relevance feedback. We again relied on a statistical language model of retrieval and used the same parameters as last year. With words as indexing terms we used queries expanded to include 60 terms and a smoothing parameter, alpha, of 0.30. When 6-grams were used instead, queries were expanded to 400 terms and alpha was set to 0.15. In both cases the top 20 documents were used as positive examples and the bottom 75 of 1000 were presumed irrelevant for the purposes of query expansion. Our official results are shown below in Table 3.

run id	topic fields	average precision	recall (at 1000)	# topics
aplmode	TD	0.4663	1792 / 1938	50
aplmoen	TD	0.3957	800 / 821	50
aplmoes	TD	0.5192	2659 / 2854	50
aplmofi	TD	0.3280	483 / 502	30
aplmofr	TD	0.4509	1364 / 1383	50
aplmoit	TD	0.4599	1039 / 1072	49
aplmonl	TD	0.5028	1773 / 1862	50
aplmosv	TD	0.4317	1155 / 1196	49

Table 3. Official results for monolingual task. The shaded row contains results for a comparable, unofficial English run.

The recall at 1000 documents is very high relative to the number of relevant documents in each of the sub-collections. Since we created runs by combining distinct runs (one using words, one using 6-grams) we should examine the individual performance using each method. Figure 1 contains a plot that shows the mean average precision obtained for each sub-collection using both approaches. We note that in English and the Romance languages, the use of words yields slightly better performance, an improvement of 0.010 to 0.025 in absolute terms. We reported observing the same trend for French and Italian during last year's evaluation [3]. In the Dutch sub-collection, little difference is seen, but 6-grams are clearly advantageous in the remaining languages. A sizeable difference is seen in German (0.035) and Swedish (0.023), and far more significantly,, in Finnish (0.13)

In Figure 1 we also plot the performance of the combined runs. Combination was generally beneficial, but due to the large disparity between n-grams and words for Finnish, the technique depressed performance in

that language compared to that which would have observed using n-grams alone. No difference due to combination was seen for German, but an improvement of between 0.016 and 0.023 was found in the remaining collections, an improvement of 3-5% in relative terms.

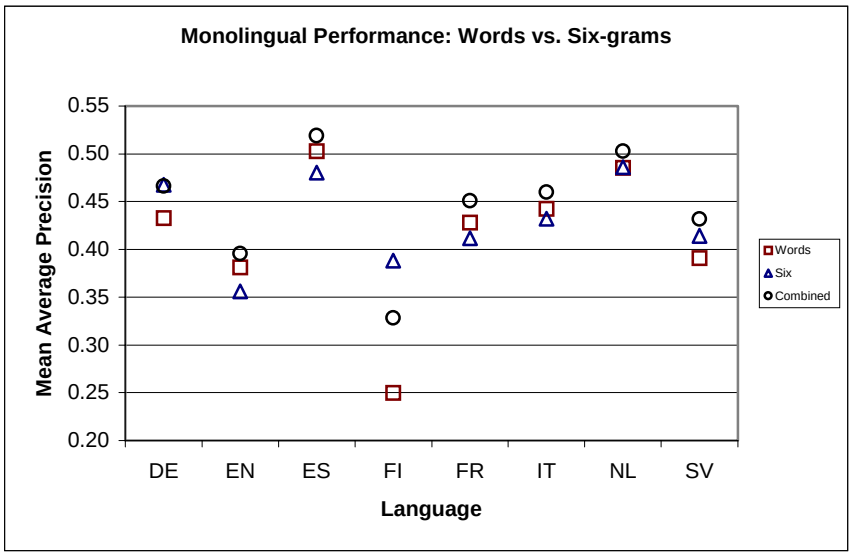

Figure 1. Comparing words and character n-grams *(n=6)* by language.

We performed the same analysis when blind relevance feedback was not performed and found similar results. There the performance was generally less than when automated feedback was performed. Also, within each language, differences between techniques were larger without feedback. By averaging across all languages, we saw that feedback improved the microaveraged mean average precision from 0.3479 to 0.4141 when words were used, and from 0.3729 to 0.4295 when 6-grams were used. If, as it seems, n-grams are more effective for retrieval in languages with complex morphology, then the fact that the two approaches achieved more similar performance when feedback was employed would support the notion that automatic relevance feedback improves performance by redressing the effect of inflectional variation.

Bilingual Experiments

Our official bilingual submissions were based on query translation when some attempt at translation was made; we submitted one run for each document collection. For each collection, save English, we created one run using the English query statements and the title and description fields. The runs are named using the template *aplbienxx*. For these 7 runs, we used pre-translation expansion using the L.A. Times collection; queries were expanded to 60 terms and we used statistical word-by-word translations mined from an aligned parallel collection. This collection is an expanded version of the corpus we obtained from the Europa web site (details follow). We used unnormalized words for these bilingual experiments because we have not yet used our parallel collection to generate statistical translations that are character n-grams – we want to investigate this, but for the evaluation, we simply used words. The final two runs, *aplbiptesa* and *aplbiptesb*, made no use of translation whatsoever. Since 10 runs were allowed to be submitted according to the track guidelines, we did submit three other runs. The first, *aplbipten*, used the Portuguese topic statements to search the English sub-collection; our motivation here was only to submit a run using these statements.

The seven runs produced using English queries first performed pre-translation expansion using the L.A. Times sub-collection. The query was expanded to include 60 words, and then each term was translated, if possible, using the Europa corpus for translation. Then two runs were made, one using pre-translation expansion alone and one using both pre- and post-translation query expansion. Scores for these two runs were normalized and merged together to form our official submission. The eighth run, *aplbipten*, used the Portuguese topic statements and no pre-translation expansion was attempted. However, two runs were still combined, one with no expansion and one that made use of blind relevance feedback. Results for these runs are shown below in Table 4.

run id	topic fields	average precision	recall (at 1000)	# topics	% mono
aplbiende	TD	0.3137	1535 / 1938	50	67.27%
aplbienes	TD	0.3602	2326 / 2854	50	69.38%
aplbienfi	TD	0.2003	388 / 502	30	61.07%
aplbienfr	TD	0.3505	1275 / 1383	50	77.73%
aplbienit	TD	0.2794	934 / 1072	49	60.75%
aplbiennl	TD	0.3516	1625 / 1862	50	69.93%
aplbiensv	TD	0.3003	1052 / 1196	49	69.56%
aplbipten	TD	0.4158	753 / 821	42	105.07%

Table 4. Official results for the bilingual task. Except for the shaded run, English was used as the source language.

The results for each run in Table 4 are not comparable to one another because a different target language collection was involved. Furthermore, the last column, which reports the comparison to a target language monolingual baseline using mean average precision, is not especially meaningful. It is unfair to compare against our monolingual baseline for two reasons. First, Voorhees has pointed out that a comparison between test-sets using different topic statements (as is the case here) is not justified even though the document collections are the same[5]; the various translations of each topic may result in queries that are significantly easier in one language than another. Second, slightly different algorithms were used in our monolingual and bilingual results. Our monolingual runs were formed through merging n-gram and word-based runs while our bilingual results only used words. Also, the bilingual runs all used pre-translation over the English collection, which itself only contained relevant documents for 42 of the 50 topics.

Improved Translation Resource

The quality of translation resources is a critical driver for CLIR performance. Therefore, it is important to select a translation approach that ensures translation of important query terms. At our disposal we had translation software (Systran, L&H Power Translator, and various on-line services), bilingual dictionaries automatically extracted from lists on the Web, and a large parallel corpus. We investigated each of these methods in our 2001 paper and found that when only a single source was used, best performance was obtained by using the parallel collection for translation. We decided to expand the parallel collection and use it for our official runs.

The collection was obtained through a nightly crawl of the Europa web site where we targeted the Official Journal of the European Union [6]. The Journal is available in each of the E.U. languages and consists mainly of governmental topics, for example, trade and foreign relations. We had data available from December 2000 through May 2002. Though focused on European topics, the time span is 5-7 years after the CLEF-2002 document collection. So, it is possible that many proper names in 1994 and 1995 will be rarely mentioned, if at all. The Journal is published electronically in PDF format and we wanted to create an aligned collection. Rather than attempt an 11-language, multiple aligned collection, we simply wasted disk space and preformed redundant alignments. At the present we have not aligned all O(n2) pairs, but instead created n alignments between English and the other languages. We used the publicly available *pdftotext* package to extract text from the PDF, but Greek text is not supported by the software so we neglected this language[1]. Once converted to text, documents were split into pieces using conservative rules for page-breaks and paragraph breaks. Many of the documents are written in outline form, or contain large tables, so this task is not trivial. Approximately 20GB of PDF documents are involved; we find that the PDF files are approximately ten times larger than the plain text versions. Thus we have about 200 MB of text in each language that may be aligned.

Once aligned, we indexed each sub-collection using the same technique described for the CLEF-2002 document collections; in particular, unnormalized words were used as indexing terms. We relied on query term translation and extracted candidate translations as follows. First, we would take a candidate term as input and identify documents containing this word in the English subset of the aligned. Up to 5000 documents were considered; we bounded the number for reasons of efficiency and because we felt that performance was not enhanced appreciably when a greater number of documents was used. If no document contained this term, then the word itself was left untranslated. Second, we would identify the corresponding documents in the target language. Third, using a similarity metric that is similar to mutual information, we

[1] We are very interested in having Danish and Portuguese document collections added to the CLEF test set.

would extract a single potential translation using the frequency of occurrence in the whole collection and the frequency in the subset of aligned documents found that are believed to contain a mapping for the original source term.

Pre-translation Expansion

We are still analyzing the effect of query expansion on retrieval performance and will report on it in the final version of the paper.

No translation

In previous work we have shown that reasonably good retrieval between two related languages is possible, without any translation at all. Though the use of cognate matches as been known for some time (*e.g.*, [1]), we found that pre-translation expansion using a comparable expansion corpus enhances performance – in some cases, by 200-300% [4]. During last year's campaign we also noted that n-grams were almost twice as effective as words in this scenario [3]. This year, we wanted to conduct similar work that looked at a variety of language pair in comparison to our pervious work which only used English as the target language. We looked at several language pairs and hoped to see a difference in performance when this method was used between close languages. Our hypothesis is that translation-less retrieval between related languages (say the Romance group) would be more effective than when this approach was used between, say, German and Spanish.

For these runs, we did not use pre-translation expansion (though we hope to examine this in the future). We did compare performance using words and n-grams. Our two official runs for this experiment we aplbiptesa and aplbiptesb. The first used 6-grams as indexing terms while the later used words. Both urns used blind relevance feedback. Results for these two runs are shown below in Table 5.

run id	topic fields	type	average precision	recall (at 1000)	# topics	% mono	%Eng bilingual
aplbiptesa	TD	6-grams	0.3325	2071 / 2854	50	64.04%	92.31%
aplbiptesb	TD	words	0.2000	1589 / 2854	50	38.52%	55.52%

Table 5. Official results for the bilingual task using no translation, the Portuguese topic statements, and the Spanish news collection.

It is interesting to note that with no translation whatsoever and the use of 6-grams as indexing terms, performance was 92% of that when English topics where translated to Spanish. This is still not a fair comparision (the English topics might be particularly hard, for example), but, it is surprisingly good. The mean precision at 5 docs for aplbiptesa was 0.3920; on average, two out of the five top documents were relevant, despite not translating the queries. We examined several other language pairs as well, but have not looked at all n(n-1) cases. These other results were not official runs.

run id	topic fields	type	average precision	recall (at 1000)	# topics	% mono
aplbideesa	TD	6-grams	0.1935	1109 / 2854	50	37.27%
aplbideesb	TD	words	0.2338	951 / 2854	50	45.03%
aplbifiesa	TD	6-grams	0.1731	1244 / 2854	50	33.34%
aplbifiesb	TD	words	0.1450	837 / 2854	50	27.93%
aplbinldea	TD	6-grams	0.2764	1025 / 1938	50	59.27%
aplbinldeb	TD	words	0.1523	610 / 1938	50	32.66%
aplbidenla	TD	6-grams	0.2440	739 / 1862	50	48.53%
aplbidenlb	TD	words	0.2444	659 / 1862	50	48.61%
aplbisvita	TD	6-grams	0.2216	614 / 1072	49	48.18%
aplbisvitb	TD	words	0.1867	302 / 1072	49	40.60%

Table 6. Results using no translation between other language pairs (languages are encoded in the run ids).

As would be expected, retrieval without translation is more effective in closely related language pairs. In the table above, we see that German retrieval against Dutch is almost 50% as effective as monolingual Dutch retrieval when using 6-grams; similarly, Dutch retrieval against German is about 60% as effective as monolingual German retrieval. This strongly suggests that for language pairs with few direct translation

resources, translation to a closely related language for which translation is feasible from the source language, can result in good cross-language retrieval performance. It remains to examine whether small length n-grams result in even greater performance, and whether pre-translation expansion improves this approach. Our previous experiments on the CLEF-2001 collection would suggest the later, but in those we only examined language pairs where English was the target language.

Multilingual Experiments

To date, our experiments in multilingual merging have not found a technique that results in producing a high quality, single ranked list from documents in many languages. Last year we experimented with methods that tried to normalize document similarity scores and to produce a single list. This year we submitted two official runs that used either merge-by-score (*aplmuena*) or merge-by-rank (*aplmuenb*). As in the past, we found these two methods comparable, but not tremendously effective. However, no more suitable method has been proposed.

We have been intrigued by work by researchers at the University of California at Berkeley that address this problem in a way that does not require score normalization. Gey et al., create a single inverted file from documents in many languages and then, to score documents, they create a composite query composed of a query statement in a single language concatenated with translations of that query in the other collection languages [2]. This approach results in a single ranked list, and it appears to work well with Berkeley's logisitic regression approach to retrieval. In the CLEF-2001 campaign we examined this method using both unnormalized words and character 5-grams. Our results with simple words were disappointing and the 5-grams, though significantly better, did not perform as well as simple merging approaches. We do not yet understand why our results are different than those reported by Berkeley, but the fact that we use a different model of retrieval may be responsible.

This year, we also attempted a dual solution to the approach described above. Rather than translate queries into every language, we created an index that contained a document that was transformed into a single language. We picked English as our interlingua and mapped each document into English using a bag-of-words approach to translation. Strictly speaking, we did not perform translation of the documents. Rather, we took the indexed document representations from our monolingual indexes, loaded a hash-table into memory that contained a bilingual wordlist, and created a new inverted-file where the posting lists were English words (or untranslatable foreign terms) that refered to documents from the different languages. We also included the native English documents. Because we felt lexical coverage was most important, we translated the documentation representations by mapping each source word into all of its candidate (English) translations. We probably should have removed stopwords, but did not do so. This process is linear in the size of the collection since the hash-table lookups are O(1) per word occurrence.

This approach creates an index with several peculiar characteristics. First, it makes the foreign language document representations a bit larger, since on average, a term may have 2 or 3 potential translations. Also, the original English documents are somewhat more focused since they don't have erroneous translations in their representations. Still, we are left with an approach where we can take a query in our preferred language (preferred here because we have good resources for it) and simply run it against our transformed document collection. This approach (*aplmuend*) appears to be 18% more effective than our officially submitted runs using normalization and merging. Interestingly, precision at a small number of documents was greatly enhanced, and recall at 1000 docs suffered; however, a subsequent combination with run *aplmuena* restored the overall recall (*aplmuenq*). Furthermore, this method creates a composite 'English' index in time linear with the collection size and requires no query-time translation or post-retrieval processing (*e.g.*, merging). See Table 7 for a comparison of this and our two official runs.

run id	topic fields	average precision	recall (at 1000)	precision at 5 docs	remarks
aplmuena	TD	0.2070	4729 / 8068	0.4680	official; score-based merge
aplmuenb	TD	0.2082	4660 / 8068	0.4480	official; rank-based merge
aplmuend	TD	0.2447	3394 / 8068	0.5760	translation of document representations
aplmuenq	TD	0.2456	4766 / 8068	0.5600	combination of aplmuena and aplmuend
aplmuenz	TD	0.2265	4772 / 8068	0.4840	score-based merge using monolingual runs

Table 7. Multilingual results.

One final thing we did for this year's mutlingual task was to try and isolate the effect of losses due to query translation and multi-collection merging. What we did was to take monolingual runs for each of the collections and attempt to merge them (*aplmuenz*). We found slightly better average precision when doing this, as might be expected. We think this is an interesting way to investigate the multilingual problem; it reduced the problem to that finding a good merging strategy, which still seems like one of the most viable approaches to MLIR.

Conclusions

We set out to investigate how well a simplified approach to CLIR would work. By applying our language-neutral philosophy, we were able to submit monolingual and bilingual runs for each of the document collections. We repeated previous experiments and confirmed that character n-grams work well in many languages, including Finnish and Swedish which we had not previously studied. N-grams appear to have a decided advantage over words in Finnish retrieval. We also examined retrieval using cognate matches between close, and less close language pairs; as expected, performance is higher (relative to a monolingual baseline) with related pairs. Finally, we implemented a novel approach to multilingual retrieval that is similar to document translation – we transformed a bag-of-words representation of documents in many languages into a corresponding set of English terms using a bilingual dictionary. This processing is efficient and can be done at indexing time. As a result, multilingual queries from a single interlingua can be processed with no additional query-time processing beyond that normal for monolingual retrieval. Our preliminary results indicate that this approach is also 18% more effective than a baseline using score normalization and merging.

References

[1] 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.

[2] F. Gey, H. Jiang, A. Chen, and R. Larson, 'Manual Queries and Machine Translation in Cross-language Retrieval and Interactive Retrieval with Cheshire II at TREC-7'. In E. M. Voorhees and D. K. Harman, eds., *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pp. 527-540, 1999.

[3] P. McNamee and J. Mayfield, 'JHU/APL Experiments at CLEF: Translation Resources and Score Normalization'..In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck (eds.), *Evaluation of Cross-Language Information Retrieval Systems: Proceedings of the CLEF 2001 Workshop, Lecture Notes in Computer Science 2406*, Springer, pp. 193-208, 2001.

[4] Paul McNamee and James Mayfield, 'Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources'. In the Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR-2002), Tampere, Finland, August 2002.

[5] E. M. Voorhees, 'The Philosophy of Information Retrieval Evaluation.' ..In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck (eds.), *Evaluation of Cross-Language Information Retrieval Systems: Proceedings of the CLEF 2001 Workshop, Lecture Notes in Computer Science 2406*, Springer, pp. 355-370, 2001.

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