JHU/APL Experiments at CLEF: Translation Resources and Score Normalization

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

The Johns Hopkins University Applied Physics Laboratory participated in three of the five tasks of the CLEF-2001 evaluation, monolingual retrieval, bilingual retrieval, and multilingual retrieval. In this paper we describe the fundamental methods we used and we present initial results from three experiments. The first investigation examines whether residual inverse document frequency can improve the term weighting methods used with a linguistically-motivated probabilistic model. The second experiment attempts to assess the benefit of various translation resources for cross-language retrieval. Our last effort is to improve cross-collection score normalization, a task essential for the multilingual problem.

Introduction

The Hopkins Automated Information Retriever for Combing Unstructured Text (HAIRCUT) is a research retrieval system developed at the Johns Hopkins University Applied Physics Laboratory (APL). The design of HAIRCUT was influenced by a desire to compare various methods for lexical analysis and tokenization; thus the system has no commitment to any particular method. With western European languages we typically use both unstemmed words and overlapping character n-grams as indexing terms, and previous experimentation has led us to believe that a combination of both approaches enhances performance [7].

We participated in three tasks at this year’s workshop, monolingual, cross-language, and multilingual retrieval. All of our official submissions were automated runs and our official cross-language runs relied on query translation using one of two machine translation systems. In the sections that follow, we first describe our standard methodology and we then present initial results from three experiments. The first investigation examines whether residual inverse document frequency can improve the term weighting methods used with a linguistically-motivated probabilistic model. The second experiment attempts to assess the benefit of various translation resources for cross-language retrieval. Our last effort is to improve cross-collection score normalization, a task essential for the multilingual problem.

Methodology

For the monolingual tasks we used twelve indices, a word and an n-gram (n=6) index for each of the six languages. For the bilingual and multilingual tasks we used the same indices with translated topic statements. Information about each index is provided in Table 1.

	# docs	collection size (MB gzipped)	name	# terms	index size (MB)
Dutch	190,604	203	words	692,745	162
			6-grams	4,154,405	1144
English	110,282	163	words	235,710	99
			6-grams	3,118,973	901
French	87,191	93	words	479,682	84
			6-grams	2,966,390	554
German	225,371	207	words	1,670,316	254
			6-grams	5,028,002	1387
Italian	108,578	108	words	1,323,283	146
			6-grams	3,333,537	694
Spanish	215,737	185	words	382,664	150
			6-grams	3,339,343	1101

Table 1. Index statistics for the CLEF-2001 test collection

Index Construction

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 CLEF-2001 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 explored language-neutral retrieval and avoided the use of stopword lists, lexicons, decompounders, stemmers, lists of phrases, or manually-built thesauri [6]. Such resources are seldom in a standard format, may be of varying quality, and worst of all, necessitate additional software development to utilize. Although we are open to the possibility that such linguistic resources may improve retrieval performance, we are interested in how far we can push performance without them. We followed the same approach this year.

We conducted our work on four Sun Microsystems workstations that are shared with about 30 other researchers. Each machine has at least 1GB of physical memory and we have access to dedicated disk space of about 200GB. The use of character n-grams increases the size of both dictionaries and inverted files, typically by a factor of five or six, over those of comparable word-based indices. Furthermore, when we use pseudo-relevance feedback we use a large number of expansion n-grams. As a consequence, runtime performance became an issue that we needed to address. Over the last year we made a number of improvements to HAIRCUT to reduce the impact of large data structures, and to allow the system to run in less memory-rich environments.

To minimize the memory consumption needed for a dictionary in a large term-space, we developed a multi-tiered cache backed by a B-tree. If sufficient memory is available, term/term-id pairs are stored in a hash table; if the hash table grows too large, entries are removed from the table, but still stored in memory as compressed B-tree nodes; if the system then runs out of memory data are written to disk.

To reduce the size of our inverted files we applied gamma compression [9] and saw our disk usage shrink to about 1/3 of its former size. HAIRCUT also generates dual files, an analogous structure to inverted files that are document-referenced vectors of terms; the dual files also compressed rather nicely.

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

Once again our approach to monolingual retrieval focused on language-independent methods. We submitted two official runs for each target language, one using the mandated <title> and <desc> fields (TD runs) and one that added the <narr> field as well (TDN runs), for a total of 10 submissions. These official runs were automated runs formed by combining results from two base runs, one using words and one using n-grams.

In all our experiments we used a linguistically motivated probabilistic model. This model has been described in a report by Hiemstra and de Vries [5], which compares the method to traditional models. This is essentially the same approach that was used by BBN in TREC-7 [8] which was billed as a Hidden Markov Model. The similarity calculation that is performed is:

$$Sim(q,d) = \prod_{t=terms} \left(a \cdot f(t,d) + (1-a) \cdot mrdf(t)\right)^{f(t,q)}$$

Equation 1. Similarity calculation.

where *f(t,d)* is the relative frequency of term *t* in document *d* (or query *q*) and *mrdf(t)* denotes the mean relative document frequency of *t*. The parameter α is a tunable parameter that can be used to ascribe a degree of importance to a term. For our baseline system we simply fix the value of α at 0.3 when words are used as indexing terms. Since individual n-grams tend to have a lower semantic value than words a lower α is indicated; we use a value of 0.15 for 6-grams. In training experiments using the TREC-8 test collection we found performance remained acceptable across a wide range of values. When blind relevance feedback is applied we do not adjust this importance value, and instead just expand the initial query.

	topic fields	average precision	recall	# topics	# ≥ median	# ≥ best	# = worst
aplmodea	TDN	0.4596	2086 / 2130	49	36	11	0
aplmodeb	TD	0.4116	2060 / 2130	49	40	6	0
aplmoena	TDN	0.4896	838 / 856	47	unofficial English run		
aplmoenb	TD	0.4471	840 / 856	47	unofficial English run		
aplmoesa	TDN	0.5518	2618 / 2694	49	36	16	1
aplmoesb	TD	0.5176	2597 / 2694	49	31	6	0
aplmofra	TDN	0.4210	1202 / 1212	49	24	11	0
aplmofrb	TD	0.3919	1195 / 1212	49	19	4	2
aplmoita	TDN	0.4346	1213 / 1246	47	32	8	1
aplmoitb	TD	0.4049	1210 / 1246	47	26	6	1
aplmonla	TDN	0.4002	1167 / 1224	50	40	12	0
aplmonlb	TD	0.3497	1149 / 1224	50	37	3	0

Table 2. Official results for monolingual task. The shaded rows contain results for comparable, unofficial English runs.

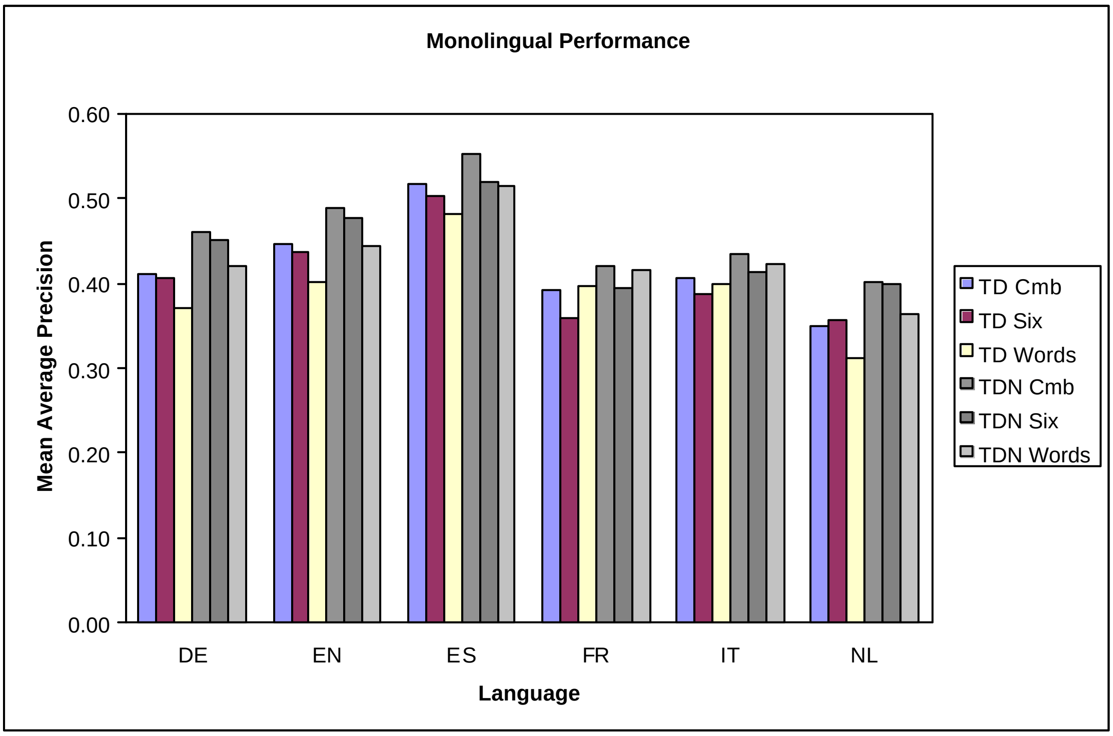

Figure 1. Comparison of retrieval performance across target languages. For each language results using both TD and TDN queries are shown when words, 6-grams, or a combination of the two is used. Unsurprisingly longer queries were more effective. 6-gram runs most often had better performance than words, but this was not the case in French or Italian. Combination of the two methods yielded a slight improvement.

We were interested in performing an experiment to see if baseline performance could be improved by adjusting the importance parameter α for each query term, Residual inverse document frequency (RIDF) [2] is a statistic that represents the burstiness of a term in the documents in which it occurs (see Equation 2 below). Terms with high RIDF tend to be distinctive, so when they are present, they occur more frequently within a document than might otherwise be expected; terms with low RIDF tend to occur indiscriminately. Numerals and adverbs, and to some extent adjectives all tend to have low RIDF. For example, the English words *briefly* and *computer* both occur in just over 5000 LA Times articles, yet computer appears 2.18 times per occurrence, on average, while briefly almost always appears just once (1.01 times on average). By taking this into account, we hope to minimize the influence that a word like briefly has on document scores (aside: Yamamoto and Church have recently published an efficient method for computing RIDF for all substrings in a collection [10]).

$$RIDF(t) = IDF(t) - \log\left(\frac{1}{1 - e^{-cf(t)}}\right)$$

Equation 2. Computing residual inverse document frequency for a term. The log term in the equation represents the expected IDF if the term had a Poisson distribution.

Our approach was as follows. For each query term, we adjust the importance value, α, for each term depending on RIDF. We linearly interpolate the RIDF value based on the minimum and maximum values in the collection and multiply by a constant k to determine the adjusted α . For these initial experiments we only considered $k=0.2$.

$$a(t) = a_{baseline} + k \cdot \frac{RIDF(t) - RIDF_{\min}}{RIDF_{\max} - RIDF_{\min}}$$

Equation 3. Computing a term-specific value for α .

We are still analyzing these results, however the preliminary indications are promising. Figure 2 shows the change in average precision when applying this rudimentary method.

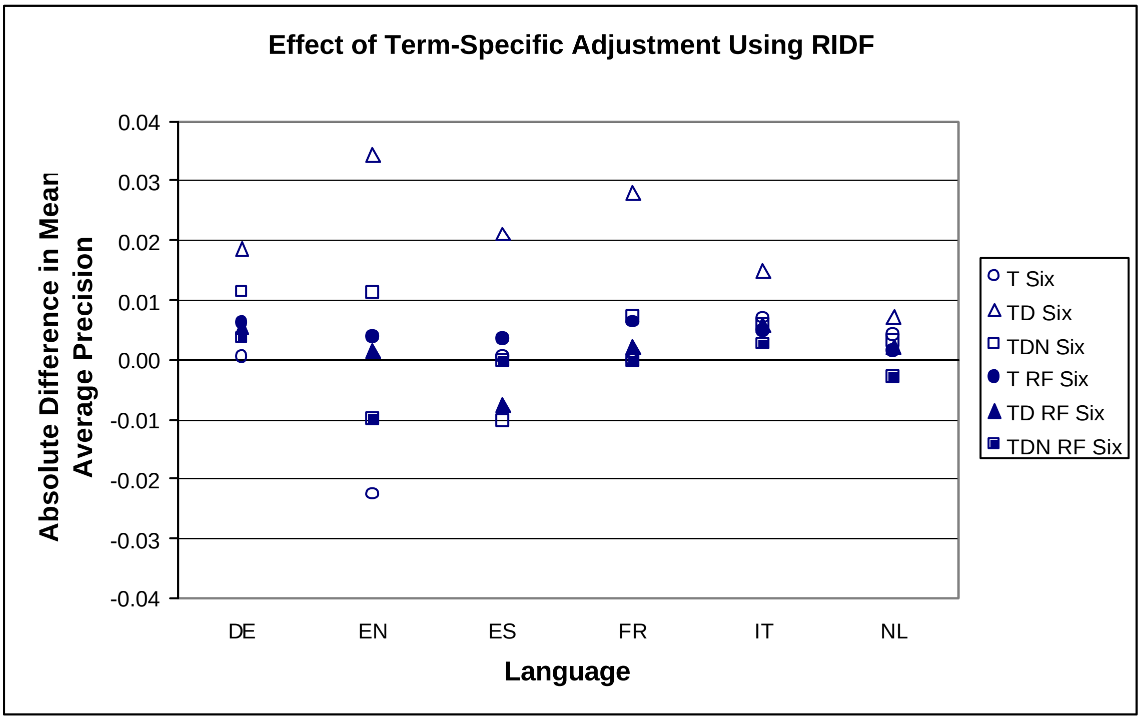

Figure 2. Impact on mean average precision when term-specific adjustments are made. 6-gram indexing shown for six query types (different topic fields and use of pseudo-relevance feedback) in each language.

We observe a small positive effect, particularly with intermediate-length queries. One possible explanation for why the improvement does not occur with very short queries (e.g., title-only) is because these queries are unlikely to contain low-RIDF terms (being short and to the point), and the adjustment in importance value is unwarranted. As yet, we have no explanation for why long queries (TDN or those with expanded queries) do not seem to gain much with this method. As time permits an analysis of individual topics may reveal what is happening.

Bilingual Experiments

Our goal for the bilingual task was to assess retrieval performance when four approaches to query translation are used, commercial machine translation software; publicly available bilingual wordlists; parallel corpora mined from the Web; and untranslated queries. The last is only likely to succeed when languages share word roots. We wanted to attempt as many of the topic languages as possible, and managed to use all but Thai.

In the past we observed good performance when commercial machine translation is used, and so all of our official runs used MT. Since only four official runs were permitted, we had a hard time choosing which topic languages to use. We attempted the Dutch bilingual task as well as the English task and ended up submitting runs using French, German, and Japanese topics against English documents, and using English topics for the Dutch documents.

	topic fields	average precision	% mono	recall	# topics	# ≥ median	# ≥ best	#= worst
aplbifren	TD	0.3519	78.7%	778 / 856	47	36	6	0
aplbideen	TD	0.4195	93.8%	835 / 856	47	31	4	2
aplbijpen	TD	0.3285	73.5%	782 / 856	47	30	3	1
aplmoenb	TD	0.4471	--	840 / 856	47	monolingual baseline		

aplbiennl	TD	0.2707	77.4%	963 / 1224	50	38	14	13
aplmonlb	TD	0.3497	--	1149 / 1224	50	monolingual baseline		

Table 3. Official results for the bilingual task

At the time of this writing we are still working on our dictionary and corpus-based methods, and will present results from these experiments in a revised version of this manuscript. We now discuss some experiments on the English bilingual collection using MT-translated and untranslated queries. Systran supports translation from Chinese, French, German, Italian, Japanese, Russian, and Spanish to (American) English; to translate Dutch, Finnish, and Swedish topics we used the on-line translator at http://www.tranexp.com/. High quality machine translation can result in excellent cross-language retrieval; our official bilingual runs achieve 81% of the performance (on average) of a comparable monolingual baseline.

Although we generally use relevance feedback and are accustomed to seeing a roughly 25% boost in performance from its use, we observed that it was not always beneficial. This was especially the case with longer queries (TDN vs. Title-only) and when the translation quality was very high for the language pair in question. In Figure 3 (below), we compare retrieval performance using words as indexing terms when relevance feedback is applied. When 6-grams were used the results were similar.

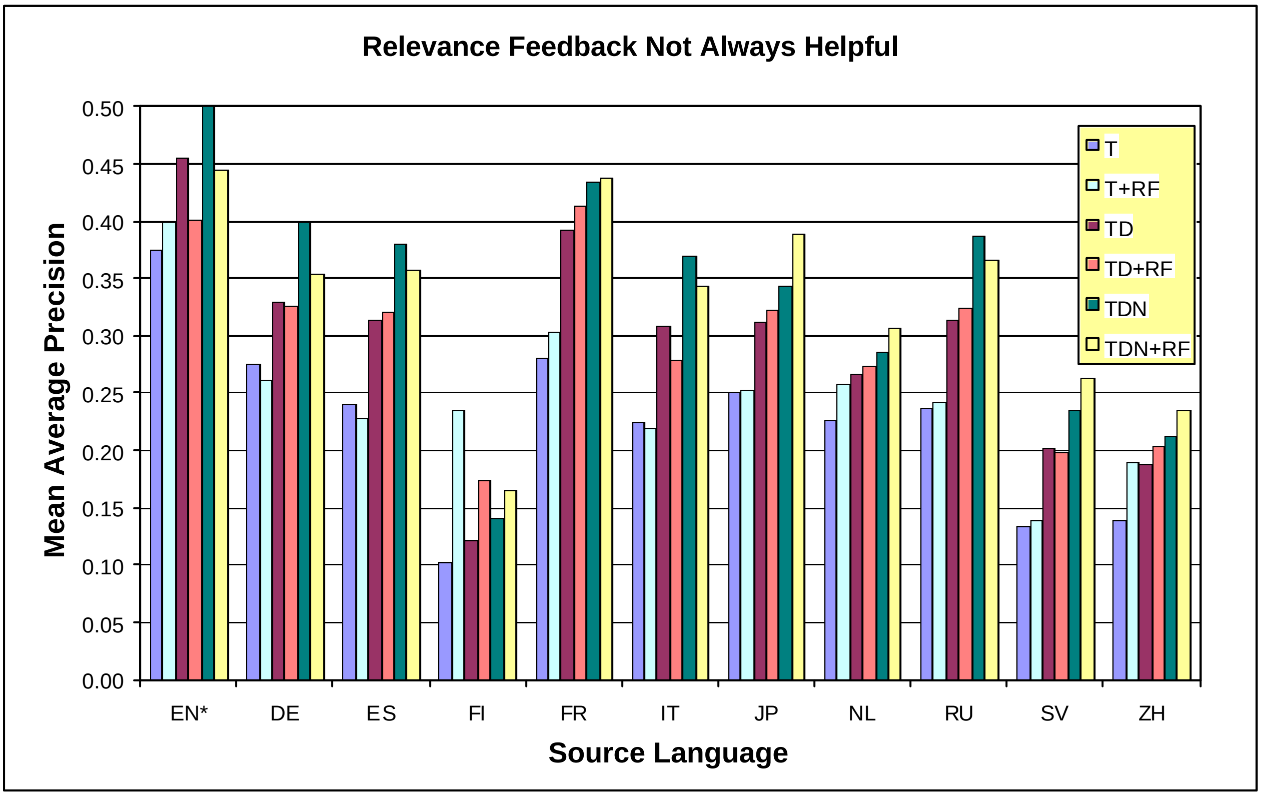

Figure 3. Bilingual performance using words as indexing terms, examining the effect of relevance feedback. Untranslated English topics are shown at the left.

Translations of Topic 41 into English

German	<DE-title> pestizide in baby food <DE-desc> reports on pestizide in baby food are looked for.
English	<EN-title> Pesticides in Baby Food <EN-desc> Find reports on pesticides in baby food.
Spanish	<ES-title> Pesticidas in foods for you drink <ES-desc> Encontrar the news on pesticidas in foods stops you drink.
Finnish	<FI-title> Suppression-compositions lasten valmisruuassa <FI-desc> Etsi raportteja suppression-aineista lasten valmisruuassa.
French	<FR-title> Of the pesticides in food for babies <FR-desc> To seek documents on the pesticides in food for babies.

Italian <IT-title> Pesticidi in the alimony for children
<IT-desc> Trova documents that they speak about the pesticidi in the alimony for children.

Japanese <JP-title>Damage by disease and pest pest control medicine in baby hood
<JP-desc>The article regarding the damage by disease and pest pest control medicine in the baby hood was searched to be.

Dutch <NL-title> Pesticide within babyvoeding
<NL-desc> Missing unpleasant documents via pesticide within babyvoeding.

Russian <RU-title> pesticides in the children's nourishment of
<RU-desc> to find articles about the pesticides in the children's nourishment of

Swedish <SV-title> Bekdmpningsmedel a baby
<SV-desc> Svk report a bekdmpningsmedel a baby.

Chinese <ZH-title> In baby food includes report which in pesticide
<ZH-desc> inquiry concerned baby food includes pesticide.

The Finnish translations are poor in quality, which explains the rather low relative performance when those topics were used. However, looking over the translated topics we observe that many untranslated terms are near cognates to the proper English word. For example, *pestizide* (German), *pesticidas* (Spanish), and *pesticidi* (Italian) are easily recognizable. Similarly, 'baby hood' is phonetically similar to 'baby food', an easy to understand mistake when Japanese phonetic characters are used to transliterate a term.

In TREC-6, Buckley et al. explored cross-language English to French retrieval using cognate matches [1]. They took an 'English is misspelled French' approach and attempted to 'correct' English terms into their proper French equivalents, projecting that 30% or so of non stopwords could be transformed automatically. Their results were unpredictably good, and they reported bilingual performance of 60% of their monolingual baseline. Although this approach is non-intuitive, it can be used as a worst-case approach when few or no translation resources are available, so long as the source and target languages are compatible. Furthermore, it can certainly be used as a lower bound on CLIR performance that can serve as a minimal standard by which to assess the added benefit of additional translation resources.

While Buckley et al. manually developed rules to spell-correct English into French, this work may be entirely unnecessary when n-gram indexing is used, since n-grams provide a form of morphological normalization. Thus we consider a more radical hypothesis than 'English is misspelled French', namely, 'other languages are English.' We now examine more closely the relative performance observed when words and 6-grams are used without spelling correction.

Figure 4 is a plot that compares the efficacy of machine-translated queries to untranslated queries for the English bilingual task. Since we have argued that relevance feedback does not have a large effect, we will only compare runs that do not use it. The data in the leftmost column is a monolingual English baseline, the unstarred columns in the central region are runs using machine translation for various source languages, and the rightmost area contains runs that used untranslated source language queries against the English collection. For each combination of translation method and source language six runs are shown using title-only, TD, or TDN topic statements and either words or 6-grams.

Several observations can be made from this plot. First, we observe that longer topic statements tend to do better than shorter ones; roughly speaking, TDN runs are about 0.05 higher than corresponding TD runs, and TD runs are about the same amount better than title-only runs. Secondly we note that 6-grams tend to outperform words; the mean relative difference among comparable MT runs is 5.95%. Looking at the various source languages we note that as a group, the Systran translated runs (DE, ES, FR, IT, JP, RU, and ZH) outperform the InterTran translated queries (FI, NL, and SV); this may reveal an underlying difference in product quality, however a better comparison would be to use languages they translate in common. Translation quality is rather poor for the Finnish and Swedish topics (InterTran) and also with the Chinese topics (Systran). Averaging across all source languages, the translated runs have performance between 41-63% of the top monolingual English run when words are used, and 41-70% when 6-grams are used.

The untranslated queries plotted on the right clearly do worse than their translated equivalents. Averaging across the seven languages encoded in ISO-8859-1, word runs achieve performance between 9-15% of the top monolingual English run, but 6-gram runs do much better and get performance between 22-34% depending on the topic fields used. The mean relative advantage when n-grams are used on these topics is 183%, almost a doubling in efficacy over words. The 6-grams achieve 54% of the performance of the machine-translated runs. Though not shown in the plot, relevance feedback actually does enhance these untranslated 6-gram runs even though we have shown that relevance feedback did not significantly affect translated topics. One final observation is that shorter queries are actually better when words are used; we suspect that this is because longer topics may contain more matching words, but not necessarily the key words for the topic.

One concern we have with this analysis is that we are comparing an aggregate measure, mean average precision. For untranslated topics, we imagine that the variance in performance is greater over many topics since some topics will have almost no cognate matches. We hope to examine individual topic behavior in the future.

We looked for this effect in other measures besides average precision. Recall at 1000 documents was effectively doubled when 6-grams were used instead of words; roughly 70% of the monolingual recall was observed. Averaged across language, Precision at 5 documents was 0.1921 when 6-grams were used with TDN topics with blind relevance feedback. Thus even this rudimentary approach can expected to find one relevant document on average in the top five documents.

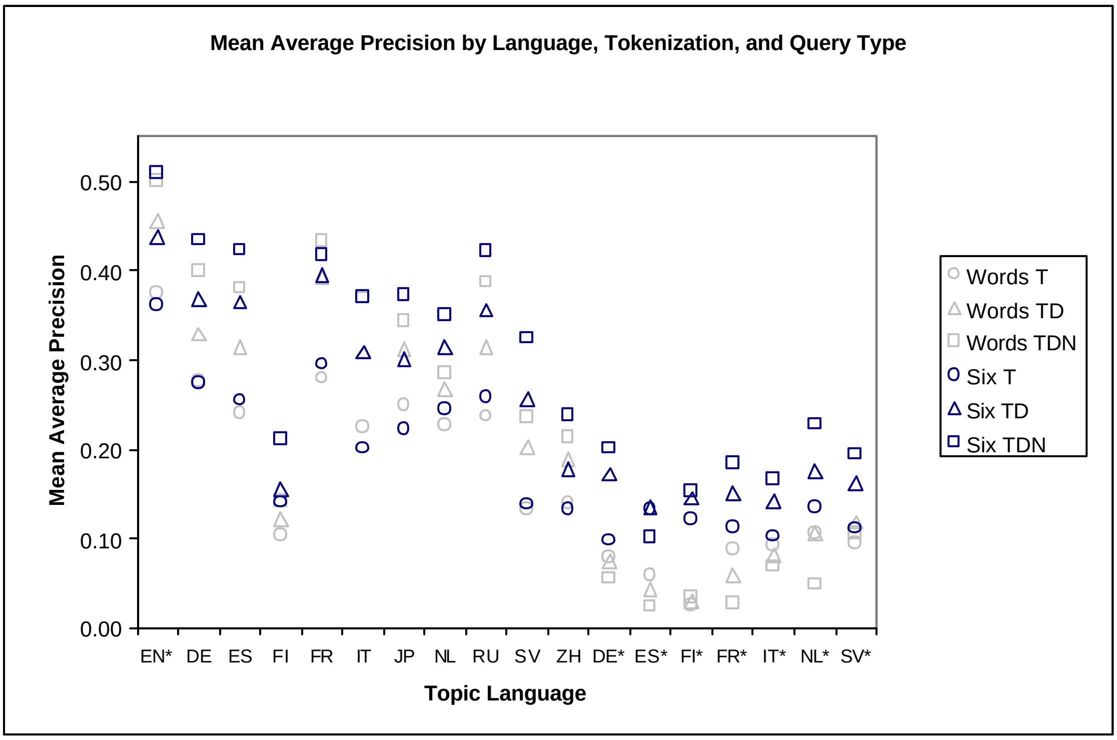

Figure 4. Comparing word and n-gram indexing on machine-translated, and untranslated topics. Untranslated topics are indicated with a star.

Multilingual Experiments

When combining several runs, one must either use document rank as a measure of the importance of a document, or try to make some sense out of system-generated scores. Using rank is problematic when the two runs cover different documents. For example, if a different index is built for each language in a multilingual retrieval task, there is no way to distinguish a language that has many relevant documents from one that has few or no relevant documents using rank alone. On the other hand, raw scores are not typically comparable. For example, the scores produced by our statistical language model are products, with one factor per query term. Even if the individual factors were somehow comparable (which they are not), there is no guarantee that a query will have the same number of terms when translated into two or more different languages. Other similarity metrics suffer from similar difficulties. Thus, score normalization is crucial if scores are to be used for run combination.

We tried a new score normalization technique this year. We viewed scores as masses, and normalized by dividing each individual score by the sum of the masses of the top 1000 documents. (Because our probabilistic calculations are typically performed in log space, and scores are therefore negative, we achieved the desired effect by using the reciprocal of a document's score as its mass.) Our previous method of score normalization was to interpolate scores for a topic within a run onto [0,1]. We were concerned that this would cause documents in languages with few or no relevant documents for a topic to appear comparable to top-ranked documents in a language with many relevant documents. While there was no appreciable difference between the two methods in this year's multilingual task (at least in average precision) we did see an eight percent improvement in precision at five documents using the new normalization (compare *aplmuena* with *aplmuend*).

We are still investigating rank-based combination as well, though we submitted no official runs using this technique. Our preliminary findings show little difference compared to score-based combination.

We were intrigued by a method that the U.C. Berkeley team used for multilingual merging in TREC-7 [4] and in last year’s CLEF workshop [3], where documents from all languages were indexed as a common collection. Queries were translated into all target languages and the resulting collective query was run against the collection. Berkeley’s results using this approach in last year’s multilingual task (run BKMUEAA1) were comparable to runs that used a merging strategy. We were inspired to try this method ourselves and built two unified indices, one using words and one using 5-grams. Using unstemmed words as indexing terms, our performance with this method was poor (run *aplmuenc*); however, we did see a significant improvement using 5-grams instead (see Table 4). Still, our attempts using a unified term space have not resulted in better scores than approaches combining separate retrievals in each target language. We will continue to examine this method because of its desirable property of not requiring cross-collection score normalization.

	topic fields	index type(s)	normalization method	average precision	recall (8138)	Prec. @ 5	# ≥ median	# ≥ best	# = worst
aplmuena	TD	words + 6-grams	mass contribution	0.2979	5739	0.5600	25	2	0
aplmuenb	TDN	words	mass contribution	0.3033	5707	0.5800	31	3	0
aplmuenc	TD	unified words	NA	0.1688	2395	0.5600	9	1	9
aplmuend	TD	words + 6-grams	linear interpolation	0.3025	5897	0.5240	32	1	0
aplmuene	TD	unified 5-grams	NA	0.2593	4079	0.5960	unofficial run		

Table 4. Multilingual results

Conclusions

The second Cross-Language Evaluation Forum workshop has offered a unique opportunity to investigate multilingual retrieval issues for European languages. We participated in three of the five tasks and were able to conduct several interesting experiments. Our first investigation into the use of term-specific adjustments using a statistical language model showed that a small improvement can be obtained when residual inverse document frequency is utilized. However, this conclusion is preliminary and we do not feel that we completely understand the mechanism involved.

Our second experiment is only partially completed; we compared bilingual retrieval performance when two query translation methods are used. The first method using extant commercial machine translation gives very good results that approach a monolingual baseline. We also showed that reasonable performance can be obtained when no attempt whatsoever is made at query translation, and we have demonstrated that overlapping character n-grams have a strong advantage over word-based retrieval in this scenario. The method is of course only practicable when related languages are involved. We think this result is significant for several reasons. First, it quantifies a lower bound for bilingual performance that other approaches may be measured against. Secondly, it implies that translation to a related language, when translation to the target language of interest is infeasible, may form the basis of a rudimentary retrieval system. We hope to augment this work by also comparing the use of parallel corpora and publicly available bilingual dictionaries in the near future.

Multilingual retrieval, where a single source language query is used to search for documents in multiple target languages, remains a critical challenge. Our attempt to improve cross-collection score normalization was not successful. We will continue to investigate this problem, which will only grow more difficult as a greater number of target languages is considered.

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] K. W. Church, 'One Term or Two?', In the *Proceedings of the 18th International Conference on Research and Development in Information Retrieval (SIGIR-95),* pp. 310-318, 1995.

[3] F. Gey, H. Jiang, V. Petras, and A. Chen, 'Cross-Language Retrieval for the CLEF Collections – Comparing Multiple Methods of Retrieval. In *Working Notes of the CLEF-2000 Workshop,* pp. 29-38, 2000.

[4] 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.

[5] D. Hiemstra and A. de Vries, 'Relating the new language models of information retrieval to the traditional retrieval models.' *CTIT Technical Report TR-CTIT-00-09,* May 2000.

[6] P. McNamee, J. Mayfield, and C. Piatko, 'A Language-Independent Approach to European Text Retrieval. In Carol Peters (ed.), *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop, Lecture Notes in Computer Science 2069,* Springer, 2001, pp 129-139, forthcoming.

[7] J. Mayfield, P. McNamee, and C. Piatko, 'The JHU/APL HAIRCUT System at TREC-8.' In E. M. Voorhees and D. K. Harman, eds., *Proceedings of the Eighth Text REtrieval Conference (TREC-8),* pp. 445-451, 2000.

[8] D. R. H. Miller, T. Leek, and R. M. Schwartz, 'A Hidden Markov Model Information Retrieval System.' In the *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR-99),* pp. 214-221, August 1999.

[9] I. Witten, A. Moffat, and T. Bell, *'Managing Gigabytes',* Chapter 3, Morgan Kaufmann, 1999.

[10] M. Yamamoto and K. Church, 'Using Suffix Arrays to Compute Term Frequency and Document Frequency for all Substrings in a Corpus'. In *Computational Linguistics,* vol 27(1), pp. 1-30, 2001.