Cross-Language Retrieval Using HAIRCUT for CLEF 2004

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

JHU/APL continued to explore the use of knowledge-light methods for scalable multilingual retrieval during the CLEF 2004 evaluation. We relied on the language-neutral techniques of character n-gram tokenization, pre-translation query expansion, statistical translation using aligned parallel corpora, fusion from disparate retrievals, and reliance on language similarity when resources are scarce. We participated in the monolingual and bilingual evaluations. Our results support the claims that n-gram based retrieval is highly effective; that fusion of multiple retrievals is helpful in bilingual retrieval; and, that reliance on language similarity in lieu of translation can outperform a high performing system using abundant translation resources and a less similar query language.

Introduction

As in the past JHU/APL's work with the HAIRCUT retrieval system for CLEF 2004 was based on language-neutral methods. In particular, we favor techniques that can be readily applied to any language or language pair. We believe that such methods are at least as effective as approaches that rely on language-specific processing, and perhaps more so. Our principal monolingual techniques include character n-gram tokenization, use of a statistical language model of retrieval, and fusion from multiple retrievals. For bilingual retrieval we focus on pre-translation query expansion using comparable collections, statistical translation from aligned parallel collections, and when translation resources are scarce, reliance on language similarity alone. We also rely on a technique that we first explored in the CLEF 2003 evaluation: direct n-gram translation, a new method of translating queries that uses n-grams rather than words as the elements to be translated [7]. This method does not suffer from certain obstacles in dictionary-based translation, such as word lemmatization, matching of multiple word expressions, and inability to handle out-of-vocabulary words such as common surnames [11].

We submitted official runs for the monolingual and bilingual tracks. For all of our runs we used the HAIRCUT system and a statistical language model similarity calculation. Some of our official runs were based solely on n-gram processing; however, we thought that by using a combination of n-grams and words or stemmed words better performance could be obtained.

Methods

HAIRCUT supports several ways of representing documents using a bag-of-terms assumption. (We emphasize that we frequently use character n-grams, not words as indexing terms.) Our general approach is to process the text of each document, reducing all terms to lower-case. Words were deemed to be white-space delimited tokens in the text; however, we preserve only the first 4 digits of a number and we truncate any particularly long tokens (those greater than 35 characters in length). We make no attempt at compound splitting. Once words are identified we optionally perform transformations on the words to create indexing terms (*e.g.,* stemming). Starting in 2003 we began removing diacritical marks, believing that they are of little importance. So-called stopwords are retained in our index and the dictionary is created from all words present in the corpus. At query time we ignore high frequency terms for reasons of run-time efficiency, and because such terms typically add little to query semantics. (By default, query terms occurring in greater than 20% of documents are ignored.)

HAIRCUT applies gamma compression to reduce the size of the inverted file, but does not store within-document positional information in the inverted index. A 'dual file', that is a document-indexed collection of term-ids and counts, is also created. Construction of this data structure doubles our on-disk space requirements, but facilitates examination of individual document representations, which is useful when generating expansion terms during pseudo relevance feedback). Our lexicon is stored as a B-tree with nodes compressed in memory to maximize the number of in-memory terms subject to physical memory limitations. For the indexes created for CLEF 2004 memory was not an issue as only O(10^6) distinct terms were found in each collection and the corresponding dictionaries were relatively small.

We continue to use a statistical language model for retrieval akin to those presented by Miller et al. [10] and Hiemstra [4] with Jelinek-Mercer smoothing[5]. In this model, relevance is defined as

$$P(D \mid Q) = \prod_{q \in Q} \left[\alpha P(q \mid D) + (1-\alpha) P(q \mid C) \right],$$

where Q is a query, D is a document, C is the collection as a whole, and α is a smoothing parameter. The probabilities on the right side of the equation are replaced by their maximum likelihood estimates when scoring a document. The language model has the advantage that term weights are mediated by the corpus. Our experience has been that this type of probabilistic model outperforms a vector-based cosine model or a binary independence model with Okapi BM25 weighting.

For the monolingual task our submitted runs were based on a combination of several base runs using different options for tokenization. JHU/APL's official bilingual submissions were based solely on stemmed words, although we had hoped to submit composite runs. Our method for combination is to normalize scores by probability mass and to then merge documents by score. All of our submitted runs were automatic runs and used only the title and description topic fields.

Monolingual Task

For our monolingual work we created several indexes for each language using the permissible document fields appropriate to each collection. We indexed the full language collection, making use of documents from 1994 and 1995, despite the fact that only half the collection was used in the evaluation. Prior to submission we discarded retrieved documents from the wrong time period. Our reasons for using the larger collection were to improve corpus statistics, pseudo relevance feedback, and for the bilingual task, pre-translation expansion. Our four basic methods for tokenization were unnormalized words, stemmed words obtained through the use of the Snowball stemmer, 4-grams, and 5-grams. We were unable to get the Snowball stemmer to work with Russian text, and we had some difficulty with it while processing Portuguese queries – many query terms were discarded. Information about each index is shown in Table 1.

Table 1. Summary information about the test collection and index data structures

language	#docs	#rel	index size (MB) / unique terms (1000s)			
			words	stems	4-grams	5-grams
EN	166754	375	143 / 302	123 / 236	504 / 166	827 / 916
FI	55344	413	90 / 978	60 / 521	136 / 138	228 / 707
FR	177450	915	129 / 328	107 / 226	393 / 159	628 / 838
PT	106821	678	101 / 303	77 / 178	292 / 152	492 / 735
RU	16715	123	26 / 253	26 / 253	44 / 136	86 / 569

Our use of 4-grams and 5-grams as indexing terms represents a departure from earlier studies using 6-grams that we justify based on recent findings [9]. The 4-grams and 5-grams seem to work equally well for monolingual retrieval. Our language model requires a single smoothing constant; we used α=0.3 with both words and stems, and α=0.8 with 4-grams and 5-grams. Each of our base runs used blind relevance feedback (queries expanded to 60 terms; terms selected using 20 top-ranked and 75 low-ranked documents). Figure 1 charts performance using our four different term indexing strategies, in isolation. The relative advantage we have previously observed n-grams to have over words is less apparent on the CLEF 2004 data.

Our official submissions were produced by fusing several base runs. We submitted three runs for each language and we report results on the English document set since the relevance judgments are available. Runs were labeled *aplmoxxa*, *aplmoxxb*, or *aplmoxxc*, where *xx* denotes the language of interest. Runs whose names end with a terminal 'a' were produced by combining a 4-gram base run with a stemmed word base run; a terminal 'b' indicates fusion of a 5-grams and stemmed words; terminal 'c' is used for runs that used both 4-grams and 5-grams. Monolingual performance based on mean average precision is reported in Table 2.

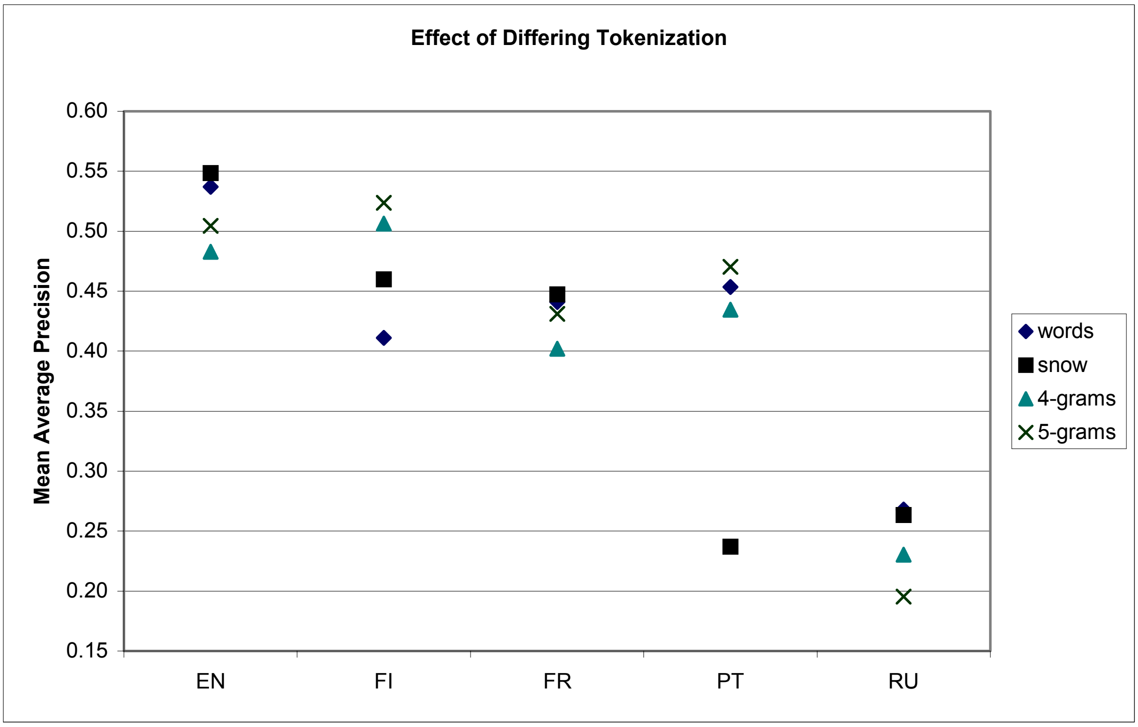

Figure 1. Relative efficacy of different tokenization methods using the CLEF 2004 test set.

Table 2. Official results for monolingual task. The shaded rows are for unofficial English runs. The maximal performing run for each language is emboldened.

run id	Fields	Terms	MAP	=Best	>=Median	Rel. Found	Relevant	# topics
aplmoena	TD	4+snow	0.5414			363	375	42
aplmoenb	TD	5+snow	0.5417			364	375	42
aplmoenc	TD	4+5	0.5070			295	375	42
aplmofia	TD	4+snow	0.5393	8	34	395	413	45
aplmofib	TD	5+snow	0.5443	6	29	394	413	45
aplmofic	TD	4+5	0.5336	8	33	392	413	45
aplmofra	TD	4+snow	0.4284	1	29	888	915	49
aplmofrb	TD	5+snow	0.4581	4	32	891	915	49
aplmofrc	TD	4+5	0.4249	2	25	810	915	49
aplmopta	TD	4+snow	0.4230	8	27	582	678	46
aplmoptb	TD	5+snow	0.4445	10	30	604	678	46
aplmoptc	TD	4+5	0.4690	11	34	589	678	46
aplmorua	TD	4+snow	0.2974	4	18	98	123	34
aplmorub	TD	5+snow	0.3076	6	19	100	123	34
aplmoruc	TD	4+5	0.2604	5	14	97	123	34

Bilingual Task

We spent a rather considerable amount of time this year in an effort to improve our translation resources. We have had consistent success using aligned parallel corpora to extract statistical translations. We have relied on this technique for single word translation; however, we recently demonstrated significant improvements in bilingual performance by translating character n-grams directly [7]. We call this 'direct n-gram translation'. Additionally we also translated stemmed words and words.

There is a consensus that lexical coverage is essential for good cross-language retrieval performance. Several studies have sought to understand the relationship between lexical coverage of translation resources and CLIR performance [2][3][7][12]. We believe that the relationship between translation coverage and performance is approximately linear. Accordingly, we sought to grow the size of our parallel collection. However, due to the nature of corpus statistics, doubling the size of a parallel collection will not necessarily double the coverage of a statistically produced translation.

For the 2002 and 2003 campaigns we relied on a single source for parallel texts, the Official Journal of the E.U. [13], which is published in the official languages (20 languages as of May 2004). The Journal is available in each of the E.U. languages and consists mainly of governmental topics, for example, trade and foreign relations. For the CLEF 2003 evaluation we had obtained 33 GB of PDF files that we distilled into approximately 300 MB of alignable text, per language. In December 2003 we began the process of mining archival issues of the Journal, beginning with 1998. This process took nearly five months. We obtained data from January 1998 through April 2004 – over six years of data. This is nearly 80 GB of PDF files, or roughly 750 MB of plain text per language. We extracted text using the *pdftotext* program; however this software cannot extract the Greek data set; we were left with data in ten languages, from which 45 possible alignments are possible. Though focused on European topics, the time span is three to ten years after the CLEF-2004 document collection. Though aware of smaller, but aligned parallel data (*e.g.,* Philip Koehn's Europarl corpus [6]) we did not utilize additional data for reasons of homogeneity and convenience.

To align data between two languages, we would:

- convert the data from PDF format to plain text (this introduced some errors, especially when processing diacritical marks in the earlier years);
- apply rules for splitting the text into sections (the data was page-aligned, we desired paragraph-sized chunks);
- and, align files using *char_align* [1].

To induce a translation for a given source language term, we proceed by:

- identifying documents (*i.e.,* approximately paragraphs) containing the source language term;
- examining the set of corresponding documents from the target language portion of the aligned collection;
- producing a score for each term that occurs in at least one of the target language paragraphs (more on this below);
- and finally, selecting the single term with the largest translation score for the source language term.

Our method for scoring candidate translations does not require translation model software such as GIZA++. Rather, we rely on information theoretic scores to rank terms. We adopt the same technique we rely on for pseudo relevance feedback – a method we have developed called *affinity sets*. Terms are weighted based on their inverse document frequency (IDF) and the difference between their relative frequency in the set of documents under consideration and the global set of documents. This measure is related to mutual information; however, we believe our technique is more general as it permits the set of documents to be identified through any means, including potentially, query-specific attempts at translation (though we do not attempt this in the experiments we report on here).

We performed pairwise alignments between languages pairs, for example, between Dutch and French. Once aligned, we indexed each pairwise-aligned collection using the technique described for the CLEF-2004 document collections. That is, we created four indexes per sub-collection, per language – one each of words, stems, 4-grams and 5-grams. This year, rather than create a translation dictionary for every term in a source language index, we translated terms on demand using the algorithm presented above. Of course, one could generate multiple translations rather than simply identifying a single one. We have not found this necessary as techniques such as pre-translation query expansion are capable of generating many terms related to a query; thus the harm introduced by a dubious translation is lessened.

We created aligned collections for the following pairs:

- Dutch and French;
- English and Finnish;
- English and French;
- English and Portuguese;
- Spanish and Finnish;

- Spanish and Portuguese;
- French and Finnish;
- and, German and French.

We had envisioned using English as a source language for the multilingual task, but not produce a submission.

At this point we should mention that the ‘proper’ translation of an n-gram is decidedly elusive concept –there is typically no single, correct answer. Nonetheless, we simply relied on the large volume of n-grams to smooth topic translation. For example, the central 5-grams of the English phrase ‘prime minister’ include ‘ime_m’, ‘me_mi’, and ‘e_min’. The derived ‘translations’ of these English 5-grams into French are ‘er_mi’, ‘_mini’, and ‘er_mi’, respectively. This seems to work as expected for the French phrase ‘premier ministre’, although the method is not foolproof. Consider n-gram translations from the phrase ‘communist party’ (parti communiste): ‘_commu’ (mmuna), ‘commu’ (munau), ‘ommun’ (munau), ‘mmuni’ (munau), ‘munis’ (munis), ‘unist’ (unist), ‘nist_’ (unist), ‘ist_p’ (ist_p), ‘st_pa’ (1_re_), ‘t_par’ (rtie_), ‘_part’ (_part), ‘party’ (rtie_), and ‘arty_’ (rtie_). The lexical coverage of translation resources is a critical factor for good CLIR performance, so the fact that almost any n-gram has a ‘translation’ should improve performance. The direct translation of n-grams may offer a solution to several key obstacles in dictionary-based translation. Word normalization is not essential since sub-word strings will be compared. Translation of multiword expressions can be approximated by translation of word-spanning n-grams. Out-of-vocabulary words, particularly proper nouns, can be partially translated by common n-gram fragments or left untranslated in close languages.

Our experience on the CLEF 2002 and 2003 bilingual tasks led us to believe that direct translation of 5-grams would likely be the most effective single technique, but that combination using runs generated by translating multiple term types would yield an improvement (see Fig. 2). It was our intent to submit such composite runs for this year’s evaluation; however, we could not complete the processing required prior to the submission deadline; it required eight indexes and runs per language pair (48 in total). Instead, we submitted runs for six language pairs using stemmed words as the sole type of token that was translated. We also submitted two runs that made no use of translation whatsoever for the language pairs Spanish to Portuguese and Bulgarian to Russian. We regret to report that we were not able to utilize the Amharic topics.

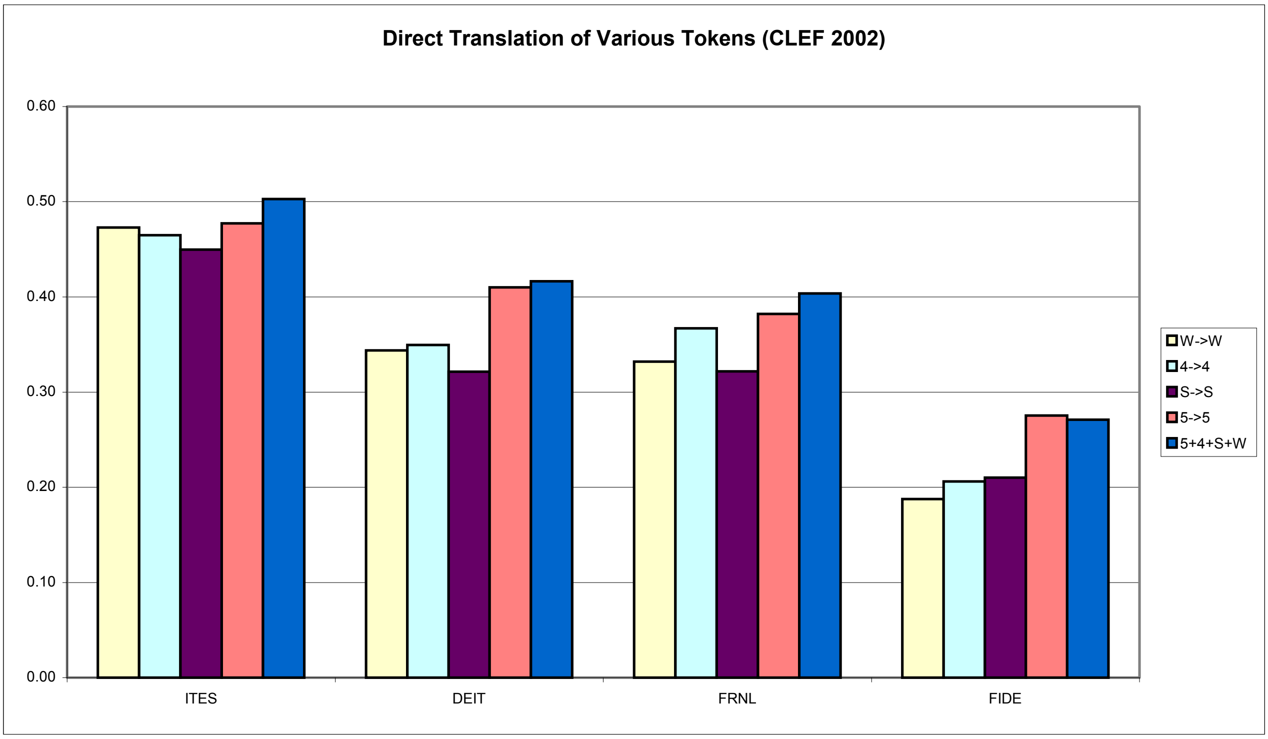

Figure 2. Relative performance of individual runs using direct translation of words, stems, and n-grams. Fusion of all four yielded the best performance in three of four cases using the CLEF 2002 bilingual test set.

Table 3. JHU/APL's official results for bilingual task.

run id	Fields	Terms	MAP	% mono	=Best	>=Median	Rel. Found	Relevant	# topics
aplbidefra	TD	4+s / s	0.3030	66.14	5	28	770	915	49
aplbienpta	TD	4+s / s	0.3414	76.91	10	23	423	678	46
aplbiesfia	TD	4+s / s	0.2982	54.79	17	36	310	413	45
aplbiespta	TD	4+s / s	0.4537	102.08	12	35	546	678	46
aplbifrfia	TD	4+s / s	0.2899	53.26	20	32	322	413	45
aplbinlfra	TD	4+s / s	0.3753	81.93	8	33	845	915	49
aplbibgrub	TD	4	0.1407	45.75	3	18	81	123	34
aplbiesptb	TD	4	0.3825	86.06	9	34	439	678	46

The performance of APL's official bilingual runs is summarized in Table 3. A terminal 'a' in the run id indicates the use of translation; a 'b' indicates no translation was attempted. The first six rows report performance against the Finnish, French, and Portuguese sub collections, using two source languages each. For these runs pre-translation expansion was incorporated by using a monolingual run based on 4-grams and stems; from these monolingual runs (against the full source language collection) 60 words were extracted. To produce our bilingual submissions, these words were stemmed and then the stems were translated into corresponding stems using parallel data for the language pair. This expanded, translated query was run against the full target language collection and retrieved documents from the wrong period were omitted.

Generally, performance for the Portuguese collection was higher than for the French and Finnish collections. We observed that translation from a very closely related language resulted in exceptional performance; for the Spanish to Portuguese run, we obtained performance 102% of a monolingual Portuguese baseline. We attribute this to the additional query expansion step that occurred (*i.e.,* pre-translation expansion). We also noted that our method of not translating queries between very closely related languages, but relying only on partial n-gram matches (*i.e.,* using 4-grams), was highly effective. This technique was so effective, that Spanish to Portuguese retrieval using 4-grams and no translation (*aplbiesptb*) outperformed translation of English queries (*aplbienpta*). Run *aplbiesptb* did at or better than median on 34 of the 46 topics. Even for language pairs with significant translation resources, language similarity should not be ignored.

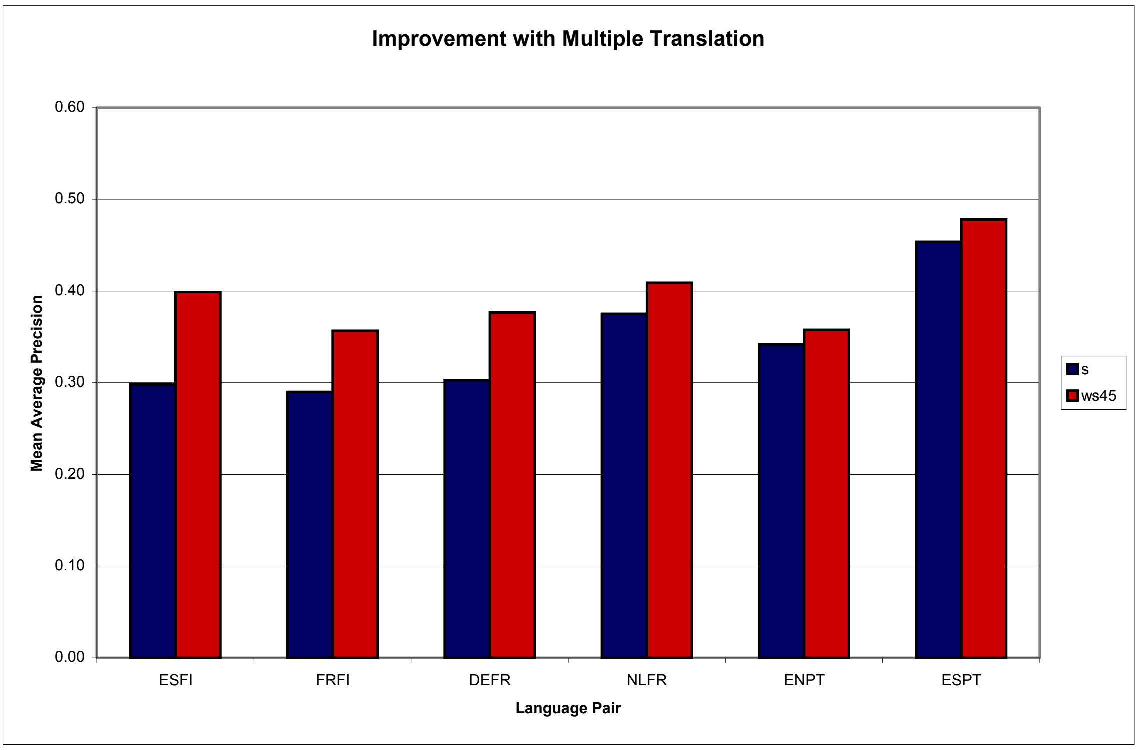

Figure 3. Improvement observed through combining multiple term translations on the CLEF 2004 Bilingual Task. The improved runs were not official submissions.

We did not have adequate opportunity to develop translation resources for Russian. Thus, we used the Bulgarian topic statements which are also in Cyrillic and hoped 'no-translation' would be effective. We report bilingual retrieval performance 45% of that of a monolingual Russian baseline, which while not as effective as between Spanish and Portuguese, might be serviceable to an end-user.

Fusion of multiple bilingual runs using translation of different token types did, in fact, confer an improvement on this year's data, as it had in previous years. Relative performance increased from between 4% and 33%, depending on the language pair, when runs using words, stems, and 4-grams and 5-grams were combined (see Fig. 3). We observed that the improvement due to this additional fusion seemed inversely proportional to the baseline monolingual performance using our official submissions.

Conclusion

JHU/APL continued its language-neutral approach to multilingual retrieval for the CLEF 2004 evaluation. For monolingual retrieval we compared words, a popular suffix stemmer, and n-grams of lengths four and five, all using the same retrieval engine and language model similarity metric. We found that n-grams continued to work well for monolingual retrieval; however, their relative efficacy compared to ordinary words appeared to be less for the CLEF 2004 data than that previously reported. We continued to combine runs produced through disparate retrievals, which we believe yields a modest improvement.

For bilingual retrieval we used direct translation of n-grams in addition to words and stems. We also found that not translating queries between closely related languages, when n-grams are used, can outperform retrieval with translation from a less similar language, even when large translation resources are available.

We will continue our work in exploring knowledge-light, language neutral approaches for retrieval. We have found the use of character n-grams, pre-translation query expansion, statistical translation using aligned parallel corpora, fusion from disparate retrievals, and reliance on language similarity when resources are scarce, all highly effective. In the future we hope to examine the identification and translation of multi-word phrases to see if such compounds can be used to improve retrieval quality.

References

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

[2] D. Demner-Fushman and D. W. Oard, 'The effect of bilingual term list size on dictionary-based cross-language information retrieval.' *Proceedings of the 36th Hawaii International Conference on System Sciences*, 2003.

[3] 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.

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

[5] F. Jelinek and R. Mercer, 'Interpolated Estimation of Markov Source Parameters from Sparse Data'. In Gelsema ES and Kanal LN eds., *Pattern Recognition in Practice*, North Holland, pp. 381-402, 1980.

[6] P. Koehn, 'Europarl: A multilingual corpus for evaluation of machine translation.' Unpublished, http://www.isi.edu/ koehn/ publications/europarl/ .

[7] P. McNamee and J. 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*, Tampere, Finland, pp. 159-166, 2002.

[8] P. McNamee and J. Mayfield, 'JHU/APL Experiments in Tokenization and Non-Word Translation.' *Working Notes of the CLEF 2003 Workshop*, pp. 19-28, 2003.

[9] P. McNamee and J. Mayfield, 'Character N-gram Tokenization for European Language Text Retrieval'. in *Information Retrieval*, 7(1-2):73-97, 2004

[10] D. Miller, T. Leek, and R. Schwartz, 'A hidden Markov model information retrieval system'. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, California, pp. 214-221, 1999.

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

[12] 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.

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