JHU/APL Experiments in Tokenization and Non-Word Translation

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

In the past we have conducted experiments that investigate the benefits and peculiarities attendant to alternative methods for tokenization, particularly overlapping character n-grams. This year we continued this line of work and report new findings reaffirming that the judicious use of n-grams can lead to performance surpassing that of word-based tokenization. In particular we examined: the relative performance of n-grams and a popular suffix stemmer; a novel form of n-gram indexing that approximates stemming and achieves fast run-time performance; various lengths of n-grams; and the use of n-grams for robust translation of queries using an aligned parallel text. For the CLEF 2003 evaluation we submitted monolingual and bilingual runs for all languages and language pairs, multilingual runs using English as a source language, and a first attempt and cross-language spoken document retrieval. Our key findings are that shorter n-grams (*n=4* and *n=5*) outperform a popular stemmer in non-Romance languages, that direct translation of n-grams is feasible using an aligned corpus, that translated 5-grams yield superior performance to words, stems, or 4-grams, and that a combination of indexing methods is best of all.

Introduction

In the past we have examined a number of issues pertaining to how documents and queries are represented. This has been a particular interest in our work with the HAIRCUT retrieval system due to the consistent success we have observed with the use of overlapping character n-grams. Simple measures that can be uniformly applied to text processing, regardless of language, reduce developer effort and appear to be at least as effective as approaches that rely on language-specific processing, and perhaps more so. They are increasingly used when linguistic resources are unavailable[11][14][15], but in general have not been widely adopted. We believe that this may be due in part to a belief that n-grams are not as effective as competing approaches (an idea that we attempt to refute here), and also due to a fear of increased index-time and run-time costs. We do not focus on the second concern here; few studies addressing the performance implications of n-gram processing have been undertaken (but see [10]), and we hope this gap is soon filled.

Over this past year we investigated several issues in tokenization. Using the CLEF 2002 and 2003 test suites as an experimental framework, we attempt to answer the following questions:

- Should diacritical marks be retained?
- What length of character n-grams results in the best performance?
- Does the optimal length vary by language?
- Are n-grams as effective as stemmed words?
- Can n-gram processing be sped up?
- What peculiarities arise when n-grams are used for bilingual retrieval?
- Are n-grams effective for cross-language spoken document retrieval?

We submitted official runs for the monolingual, bilingual, multilingual tracks and participated in the first cross-language spoken document benchmark. For all of our runs we used the HAIRCUT system and a statistical language model similarity calculation. Many of our official runs were based on n-gram processing though we found that by using a combination of n-grams and stemmed words better performance can be obtained. For our bilingual runs we relied on pre-translation query expansion. We also developed a new method of translating queries, using n-grams rather than words as the elements to be translated. This method does not suffer from several key obstacles in dictionary-based translation, such as word lemmatization, matching of multiple word expressions, and out-of-vocabulary words such as common surnames [12].

Methods

HAIRCUT supports a variety of indexing terms and represents documents using a bag-of-terms model. Our general method is to process the text for each document, reducing all terms to lower-case. Generally 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). Once words are identified we optionally perform transformations on the words to create indexing terms (e.g., stemming). So-called stopwords are retained in our index and the dictionary is created from all words present in the corpus.

We have wondered whether diacritical marks have much effect upon retrieval performance - for a long time we have been retaining diacritical marks as part of our ordinary lexical processing, in keeping with a keep-it-simple approach. One principled argument for retaining inflectional marks is that they possess a deconflationary effect when content words that differ only in diacritics have different meaning. For example, the English words resume (to continue) and résumé (a summary of one's professional life) can be distinguished by differences in diacritics. On the other hand, such marks are not always uniformly applied, and furthermore, if retained, might distinguish two semantically related words. Stephen Tomlinson investigated preservation of diacritics using the CLEF 2002 collection and reported that it was helpful in some cases (Finnish) and harmful in others (Italian and French) [16]. We found similar results (see Table 1), though the effect is seen only for words, not n-grams. As there is practically no effect, we opted to remove such accents routinely. Intuitively we thought that removing the distinction might improve corpus statistics when n-grams are used. Whenever stemming was used, words were first stemmed, and then any remaining marks were removed; this enabled the stemmer to take advantage of marks when present. N-grams were produced from the same sequence of words; however, we attempt to detect sentence boundaries to prevent generating n-grams across sentence boundaries.

language	DE	EN	ES	FI	FR	IT	NL	SV
words	-0.0002	0.0028	0.0146	-0.0363	0.0139	0.0076	-0.0005	0.0045
4-grams	-0.0028	-0.0093	0.0019	0.0075	0.0077	-0.0090	0.0009	-0.0056

Table 1. Absolute difference in mean average precision when accented marks were removed.

HAIRCUT uses gamma compression to reduce the size of the inverted file. Within-document positional information is not retained, but both document-id and term frequencies are compressed. We also produce a 'dual file' that is a document-indexed collection of term-ids and counts. Construction of this data structure doubles our on-disk space requirements, but confers advantages such as being able to quickly examine individual document representations. This is particularly useful for automated (local) query expansion. Our lexicon is stored as a B-tree but nodes are compressed in memory to maximize the number of in-memory terms subject to physical memory limitations. For the indexes created for CLEF 2003 memory was not an issue as only O(10^6) distinct terms were found in each collection.

We use a statistical language model for retrieval akin to those presented by Miller et al. [9] and Hiemstra [2] with Jelinek-Mercer smoothing[3]. 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, bilingual, and multilingual tasks, all of our submitted runs were based on a combination of several base runs. Our method for combination was to normalize scores by probability mass and to then merge documents by score. All of our runs were automatic runs and used only the title and description topic fields.

Monolingual Experiments

For our monolingual work we created several indexes for each language using the permissible document fields appropriate to each collection. Our four basic methods for tokenization were unnormalized words, stemmed words obtained through the use of the Snowball stemmer, 4-grams, and 5-grams. Information about each index is shown in Table 2.

language	#docs	%docs	#rel	%rel	index size (MB) / unique terms (1000s)			
					words	stems	4-grams	5-grams
DE	294805	18.3	1825	18.2	265 / 11.88	219 / 860	705 / 219	1109 / 1230
EN	166754	10.3	1006	10.0	143 / 302	123 / 235	504 / 166	827 / 917
ES	454041	28.2	2368	23.6	303 / 525	251 / 347	990 / 217	1538 / 1144
FI	55344	3.4	483	4.8	89 / 977	60 / 520	136 / 138	229 / 709
FR	129804	8.1	946	9.4	91 / 262	76 / 178	277 / 144	440 / 724
IT	157558	9.7	809	8.0	115 / 374	92 / 224	329 / 144	529 / 721
NL	190605	11.8	1577	15.7	161 / 683	147 / 575	469 / 191	759 / 1061
RU	16715	1.0	151	1.5	25 / 253	25 / 253	44 / 136	86 / 569
SV	142819	8.9	889	8.8	94 / 505	80 / 361	258 / 162	404 / 863
total	1608445		10054		1286 MB	1073 MB	3712 MB	5921 MB

Table 2. Summary information about the test collection and index data structures

From the table above it can be seen that the percentage of relevant documents for each subcollection is closely related to its contribution to the overall number of documents. This would suggest that collection size might be a useful factor for multilingual merging. We also note that n-gram indexing results in increased disk storage costs. This cost is driven by the increased number of postings in the inverted file when n-gram indexing is performed.

Our use of 4-grams and 5-grams as indexing terms represents a departure from previous work using 6-grams [6]. We conducted tests using various lengths of n-grams for all eight CLEF 2002 languages and found that choices of *n=4* or *n=5* performed best. Figure 1 charts performance using six different term indexing strategies; a value of *α=0.5* was used throughout and no relevance feedback was attempted.

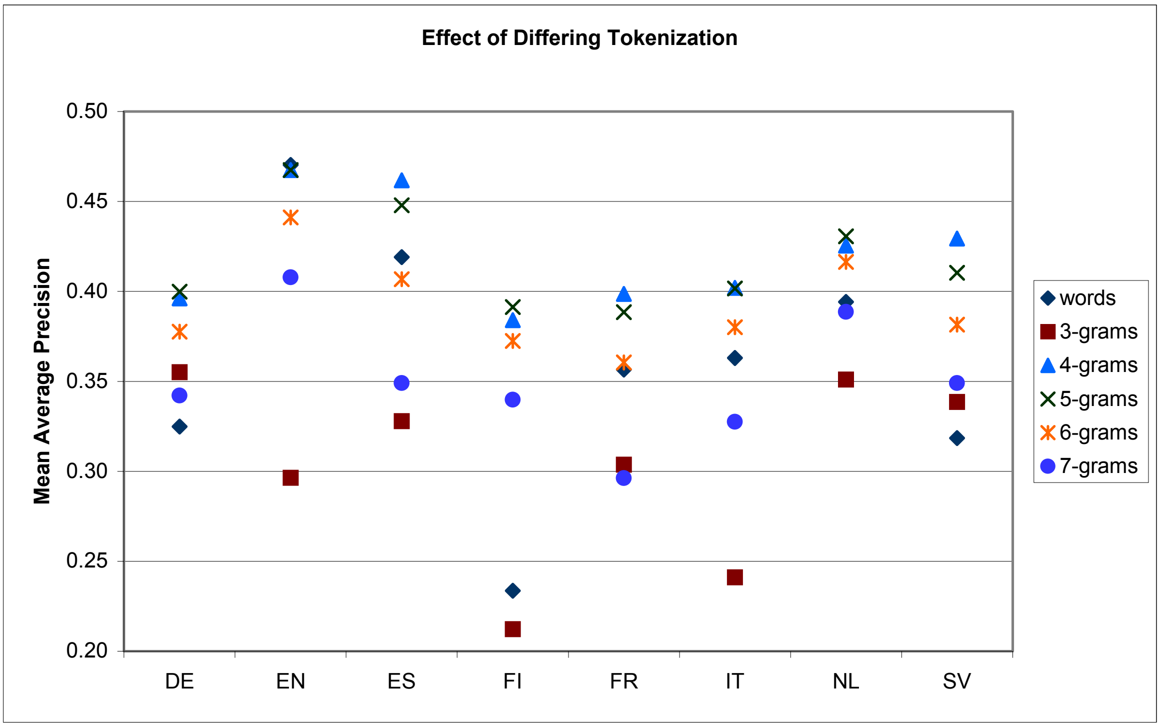

Figure 1. Relative efficacy of different tokenization methods using the CLEF 2002 test set. Note that blind relevance feedback was not used for these runs.

We determined that use of *n=4* or *n=5* is best in all eight languages though it is hard to distinguish between the two. 6-grams are clearly not as effective in these languages. There are differences in performance depending on the value of smoothing constant, α, that is used, though we have yet to test whether these differences are significant or merely represent overtraining on the 2002 test set. The effect of smoothing parameter selection in language model-based retrieval was investigated by Zhai and Lafferty [17]. We report on our results investigating the effect of n-gram length, with additional detail and further experiments in a forthcoming manuscript [8].

In additional to determining good values for *n*, we also wanted to see if n-grams remained an attractive technique in comparison to stemmed words. Having no substantive experience with stemming, we were pleased to discover that the Snowball stemmer [13], a derivative of the Porter stemmer extended to many languages by Porter, provides a set of rules for all of the CLEF 2003 languages. Furthermore, the software contains Java bindings so it fit seamlessly with the HAIRCUT system. We decided to make a comparison between raw words, stems, 4-grams, 5-grams, and a surrogate technique based on n-grams that might approximate stems. Our n-gram approximation to stemming was based on picking the word-internal n-gram for each word with lowest document frequency (i.e., we picked the least common n-gram for each word). As an example, consider the words 'juggle', 'juggles', and 'juggler'. The least common 5-gram for the first two is 'juggl', however the least common 5-gram for 'juggler' is 'ggler'[1]. The least common 4-gram for all three words is 'jugg'. We hypothesize that high IDF n-gram affixes will span portions of words that exhibit little morphological variation.

This method has the advantage of providing some morphological normalization, but it does not increase the number of postings in an inverted file. This can be viewed either as a way to approximate stems or a way of lowering the computational cost of using n-grams. We found that n-grams did outperform stems, and that our pseudo stems based on n-grams were better than raw words, but not as effective as a rule-based stemmer (see Figure 2). Details about this work can be found in Mayfield and McNamee [5].

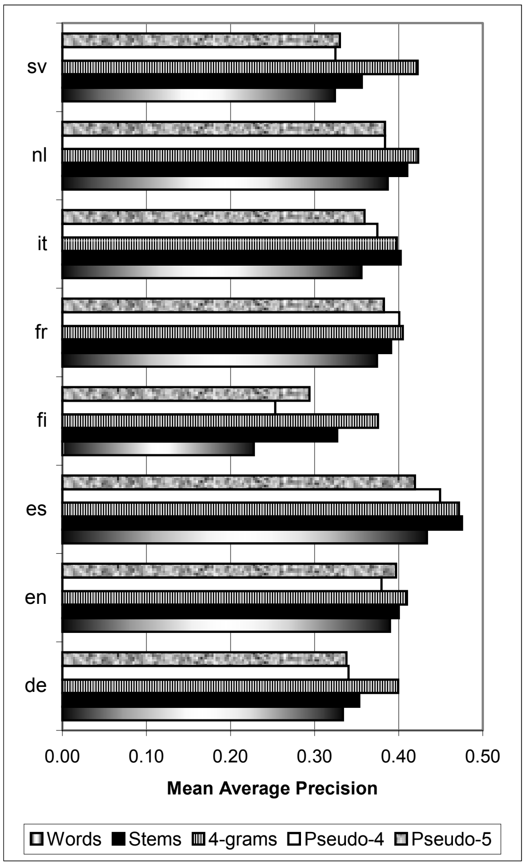

Figure 2. Comparing words, stemmed words, 4-grams, and approximate stemming (2002 collection).

[1] The Snowball stemmer also fails to transform juggler to a canonical form.

On the 2003 test collection we produced base runs for words, stems (using the Snowball stemmer), 4-grams, and 5-grams. Performance (based on average precision) for each is reported in Table 3. All of these runs used blind relevance feedback and used an α value of 0.3 with words and stems, or 0.8 with n-grams. None of these runs were submitted as official runs; instead, we created hybrid runs using multiple methods. In the past we have found that combination from multiple runs can confer a nearly 10% improvement in performance. Savoy has also reported improvements from multiple term types [15].

	DE	EN	ES	FI	FR	IT	NL	RU	SV
words	0.4175	0.4988	0.4773	0.3355	0.4590	0.4856	0.4615	0.2550	0.3189
stems	0.4604	0.4679	0.5277	0.4357	0.4780	0.5053	0.4594	0.2550	0.3698
4-grams	0.5056	0.4692	0.5011	0.5396	0.5244	0.4313	0.4974	0.3276	0.4163
5-grams	0.4869	0.4610	0.4695	0.5468	0.4895	0.4568	0.4618	0.3271	0.4137

Table 3. Mean average precision for CLEF 2003 base runs with maximal values highlighted.

To produce our official monolingual runs we decided to combine runs based on the Snowball stemmer with runs using n-grams as indexing terms. Runs named *aplmoxxa* used 4-grams and stems while runs named *aplmoxxb* used 5-grams and stems. However, due to a mistake while creating the scripts used to produce all of our runs, we inadvertently failed to perform blind relevance feedback for our monolingual submissions. Routinely we expand queries to 60 terms using additional terms ranked after examining the top 20 and bottom 75 (of 1000) documents. Failing to use blind relevance feedback had a detrimental effect on our official runs. Our official monolingual runs are described in Table 4 and corrected scores are presented on the far right.

run id	MAP	=Best	>=Median	Rel. Found	Relevant	# topics	MAP’	% change
aplmodea	0.4852	2	31	1721	1825	56	0.5210	7.39%
aplmodeb	0.4834	2	27	1732			0.5050	4.46%
aplmoena	0.4943			977	1006	54	0.5040	1.96%
aplmoenb	0.5127			980			0.5074	-1.03%
aplmoesa	0.4679	3	32	2226	2368	57	0.5311	13.50%
aplmoesb	0.4538	3	32	2215			0.5165	13.82%
aplmofia	0.5514	12	31	475	483	45	0.5571	1.03%
aplmofib	0.5459	9	31	475			0.5649	3.49%
aplmofra	0.5228	9	35	924	946	52	0.5415	3.58%
aplmofrb	0.5148	9	37	920			0.5168	0.39%
aplmoita	0.4620	7	21	776	809	51	0.4784	3.54%
aplmoitb	0.4744	8	22	771			0.4982	5.02%
aplmonla	0.4817	3	42	1485	1577	56	0.5088	5.63%
aplmonlb	0.4709	2	40	1487			0.4841	2.86%
aplmorua	0.3389	2	17	115	151	28	0.3728	10.00%
aplmorub	0.3282	4	16	113			0.3610	10.00%
aplmosva	0.4515	7	36	840	889	53	0.4358	-3.47%
aplmosvb	0.4498	6	38	838			0.4310	-4.18%

Table 4. Official results for monolingual task. The shaded row contains results for a comparable, unofficial English run. The two columns at the far right report a corrected value for mean average precision when blind relevance feedback is applied, and the relative difference compared to the corresponding official run.

It appears that several of our runs would have increased substantially if we had correctly used blind relevance feedback. Relative improvements of more than 5% were seen in German, Russian, and Spanish although performance would have dropped slightly in Swedish. The German and Spanish document collections are the two largest in the entire test suite. We wonder if relevance feedback may be more beneficial when larger collections are available, a conjecture partially explored by Kwok and Chan [4].

Bilingual Experiments

This year the Bilingual task focused on retrieval involving four language pairs, which notably did not contain English as a source or target language. This is only significant because of the difficulty in locating direct translation resources for some language pairs and the fact that many translation resources are available when

English is one of the languages involved. The four language pairs are German to Italian, Finnish to German, French to Dutch, and Italian to Spanish.
For the 2002 campaign we relied on a single translation resource: bilingual wordlists extracted from parallel corpora. We built a large alignable collection from a single source, the Official Journal of the EU [18], and we again used this resource as our only source of translations for 2003. The parallel corpus grew by about 50% this year, so a somewhat larger resource was available. First we describe the construction of the parallel corpus and the extraction of our bilingual wordlists, then we discuss our overall strategy for bilingual retrieval, and finally we report on our official results.

Our collection was obtained through a nightly crawl of the Europa web site where we targeted the Official Journal of the European Union [18]. 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 2003. Though focused on European topics, the time span is 5 to 8 years after the CLEF-2002 document collection. The Journal is published electronically in PDF format and we wanted to create an aligned collection. We started with 33.4 GB of PDF documents and converted them to plain text using the publicly available *pdftotext* software (version 1.0). 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 pre-alignment processing is not easy. We ended up with about 300MB of text, per language, that could be aligned. Alignment was carried out using the *char_align* program [1]. In this way we created an aligned collection of approximately 1.2 million passages; these 'documents' were each about 2 or 3 sentences in length.

We performed pairwise alignments between languages pairs, for example, between German and Italian. Once aligned, we indexed each pairwise-aligned collection using the technique described for the CLEF-2003 document collections. Again, we created four indexes per sub-collection, per language – one each of words, stems, 4-grams and 5-grams. Our goal was to support query term translation, so for each source language term occurring in at least 4 documents, we attempted to determine a translation of the same token type in the target language. At this point we should mention that the 'proper' translation of an n-gram is decidedly slippery – clearly there can be 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 be partially translated by common n-gram fragments or left untranslated in close languages.

We extracted candidate translations as follows. First, we would take a candidate term as input and identify documents containing this term in the source language subset of the aligned collection. 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 it was left untranslated. Second, we would identify the corresponding documents in the target language. Third, using a statistic that is similar to mutual information, we would extract a single potential translation. Our statistic is a function of the frequency of occurrence in the whole collection and the frequency in the subset of aligned documents. In this way we extracted the single-best target language term for each source language term in our lexicon (not just the query terms in the CLEF topics). When 5-grams were used this process took several days.

Table 5 lists examples of translating within the designated language pairs using each type of tokenization. Mistakes are evident; however, especially when pre-translation expansion is used the overall effectiveness is quite high. We believe the redundancy afforded by translating multiple n-grams for each query word also reduces loss due to erroneous translations. Finally, incorrect translations may still prove helpful if they are a collocation rather than an actual translation.

	Desired Mapping	DEIT		FIDE		FRNL		ITES	
		DE	IT	FI	DE	FR	NL	IT	ES
words	milk	milch	latte	maidon	milch	lait	melk	latte	leche
	olympic	olympische	olimpico	olympialaisiin	olympischen	olympique	olympisch	olimpico	olimpico
stems	milk	milch	latt	maido	milch	lait	melk	latt	lech
	olympic	olymp	olimp	olymp	olymp	olymp	olympisch	olimp	olimp
4-grams	first 4-gram (milk)	milc	latt	maid	land	lait	melk	latt	lech
	last 4-gram (milk)	ilch	latt	idon	milc	lait	melk	atte	acte
	first 4-gram (olympic)	olym	olim	olym	olym	olym	olym	olim	olim
	last 4-gram (olympic)	sche	rope	siin	n_au	ique	isch	pico	pico
5-grams	first 5-gram (milk)	milch	_latt	maido	milch	_lait	_melk	latte	leche
	last 5-gram (milk)	milch	_latt	aidon	milch	lait_	_melk	latte	leche
	first 5-gram (olympic)	olymp	olimp	olymp	olymp	olymp	_olym	olimp	olimp
	last 5-gram (olympic)	ische	urope	isiin	ichen	pique	pisch	mpico	_olim

Table 5. Examples of term-to-term translation

We remain convinced that pre-translation query expansion is a tremendously effective method to improve bilingual performance. Therefore we used each CLEF 2003 document collection as an expansion collection for the source language queries. Queries were expanded to a list of 60 terms, and then we attempted to translate each using our corpus-derived resource. In the past we have been interested in using n-grams as terms, however, we have worked with bilingual wordlists for translation. This year we decided to create translingual mappings using the same tokenization in both the source and target languages. Thus for each of the four language pairs, we created four different lists (for a total of 16): one list per type of indexing term (*i.e.,* word, stem, 4-gram, or 5-gram). Again using experiments on the CLEF 2002 collection, we determined that mappings between n-grams were more efficacious than use of word-to-word or stem-to-stem mappings. Thus different tokenization can be used for initial search, pre-translation expansion, query translation, and target language retrieval. In testing we found the best results using both n-grams and stems for an initial source-language search, then we extracted ordinary words as 'expansion' terms, and finally we translated each n-gram contained in the expanded source language word list into n-grams in the target language (or stems into stems, as appropriate). The process is depicted in Figure 3:

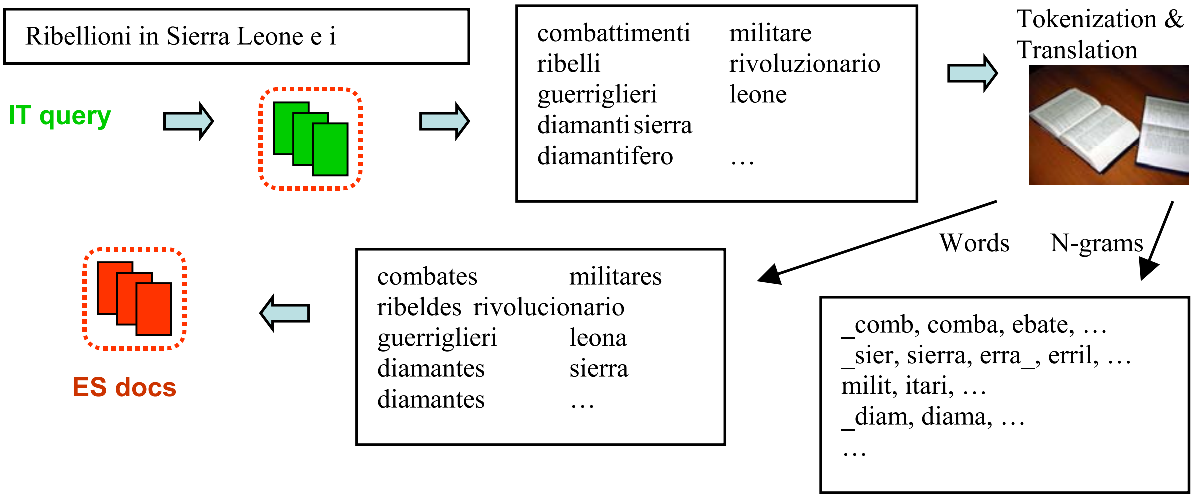

Figure 3. Illustration of bilingual processing. The initial input to translation is an expanded list of plain words extracted from a set of documents obtained by retrieval in the source language collection. These words are optionally tokenized (*e.g.*, to stems or n-grams), and the constituent query terms are then translated using the mappings derived from the parallel texts. Multiple base runs are combined to create a final ranked list.

The performance of APL's official bilingual runs are described in Table 6.

run id	MAP	% mono	=Best	>=Median	Rel. Found	Relevant	# topics
aplbideita	0.4264	89.88	11	38	789	809	51
aplbideitb	0.4603	97.03	12	45	780		
aplbifidea	0.3454	71.19	16	39	1554	1825	56
aplbifideb	0.3430	70.69	16	42	1504		
aplbifrnla	0.4045	83.97	15	33	1493	1577	56
aplbifrnlb	0.4365	90.62	13	33	1442		
aplbiitesa	0.4242	90.66	5	32	2174	2368	57
aplbiitesb	0.4261	91.07	4	38	2189		

Table 6. Official results for bilingual task.

Our runs named *aplbixxyy**a*** are bilingual runs that were translated directly from the source language to the target language; each run was a combination of four base runs that either used words, stems, 4-grams, or 5-grams, with (post-translation) relevance feedback. The runs named *aplbixxyy**b*** were combined in the same way, however the four constituent base runs did not make use of post-translation feedback. When words or stems were used a value of 0.3 was used for alpha; when n-grams were used the value was 0.5. The base runs are compared in Figure 4.

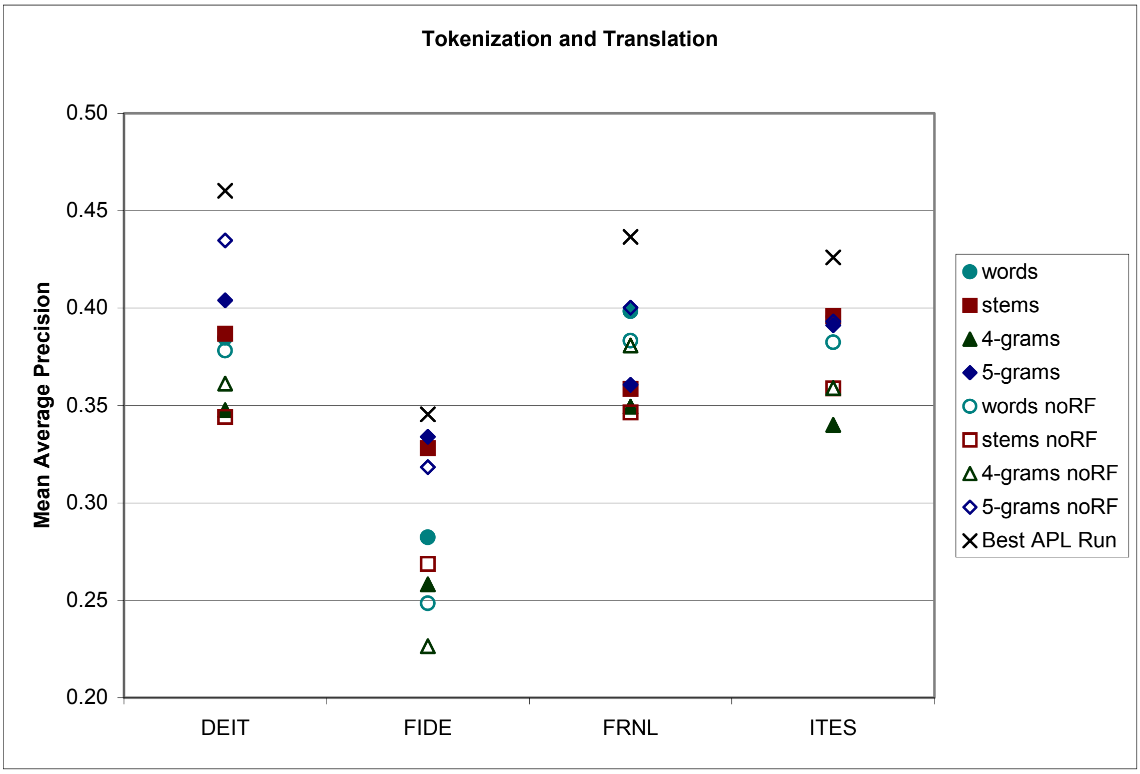

Figure 4. Analysis of the base-runs used for bilingual retrieval. The best APL run was achieved in each instance through run combination.

From observing the data in Table 6 and Figure 4, it would appear that the use of post-translation feedback did not enhance performance when multiple runs were combined. The two types of runs seemed to perform similarly in two language pairs (Finnish to German and Italian to Spanish); however, the merged runs without relevance feedback did better for the German to Italian and French to Dutch runs.

Combination of methods resulted in between a 3 and 10% gain depending on language pair. We have not yet had the opportunity to retrospectively analyze the contribution to our overall performance of pre-translation expansion.

Multilingual Experiments

We initially thought to create runs for the multilingual task in the exact same way as for the bilingual task. However, we decided to use English as our source language and we had to create translation lists for seven languages using four tokenization types (a total of 28 mappings). Construction of the 5-gram lists took longer than expected and so we had to modify our plans for our official submission. We decided to submit a hybrid run based on words, stems, and 4-grams; merging was again accomplished using normalized scores. As with the bilingual task, runs ending in 'a' denote the use of post-translation relevance feedback, while runs ending in 'b' did not use feedback (see Table 7).

run id	Task	MAP	=Best	>=Median	Rel. Found	Relevant	# topics
aplmuen4a	4	0.2926	3	33	4377	6145	60
aplmuen4b	4	0.2747	0	34	4419		
aplmuen8a	8	0.2377	4	28	5939	9902	60
aplmuen8b	8	0.2406	1	41	5820		

Table 7. APL results for multilingual task.

Spoken Document Evaluation

This was our first time using the TREC-8 and TREC-9 spoken document dataset. Our submissions were created in very short order – in one day. We pre-processed the data so it had similar SGML markup as the *ad hoc* TREC collections and then indexed the English text using only 5-grams. The index took 33 minutes to build. We did not make use of any collection expansion for these runs. Our processing was similar to the work we did for the bilingual track, except that we used only 5-grams as translation terms and did not use pre-translation expansion (which was not permitted for 'primary' submissions).

The runs we submitted for the spoken document evaluation are summarized in Table 8.

run id		Task / Condition	MAP
aplspenena	EN	Monolingual	0.3184
aplspfrena	FR	Primary	0.1904
aplspdeena	DE	Primary	0.2206
aplspnlena	NL	Secondary	0.2269
aplspitena	IT	Secondary	0.2046
aplspesena	ES	Secondary	0.2395

Table 8. Submissions for the Cross-Language Spoken Document Evaluation

Conclusions

For the first time we were able to directly compare words, various lengths of character n-grams, a suffix stemmer, and an n-gram alternative to stemming, all using the same retrieval engine. We found that n-grams of shorter lengths (*n=4* or *n=5*) were preferable across the CLEF 2003 languages and that n-grams generally outperformed use of the Snowball stemmer: 4-grams had a 8% mean relative advantage across the 9 languages compared to stems; however stemming was better in Italian and Spanish (by 17% and 5% respectively). We found best performance can be obtained using a combination of methods. If emphasis is placed on accuracy over storage requirements or response time, this approach is reasonable. For bilingual retrieval we identified a method for direct translation of n-grams instead of word-based translation. Without the use of relevance feedback, 5-grams outperformed stems by an average of 17% over the four bilingual pairs though 4-grams appeared to lose much of their monolingual superiority. When feedback was used, the gap narrowed substantially.

This work should not be taken as an argument against language resources, but rather as further evidence that knowledge-light methods can be quite effective, when optimized. We are particularly excited about the use of non-word translation (i.e., using direct n-gram translation) as this appears to have the potential to avoid several pitfalls that plague dictionary-based translation of words.

We are still analyzing our results from the multilingual and spoken-document tracks and hope to report on them more fully in our revised paper.

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

[3] 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.

[4] 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.

[5] J. Mayfield and P. McNamee, 'Single N-gram Stemming', To appear in the *Proceedings of the Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

[6] P. McNamee and J. Mayfield, 'Scalable Multilingual Information Access'. To appear in the *Proceedings of the CLEF 2002 Workshop*.

[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, 'Character N-gram Tokenization for European Language Text Retrieval'. To appear in *Information Retrieval*.

[9] 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.

[10] E. Miller, D. Shen, J. Liu, and C. Nicholas, 'Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System.' In the *Journal of Digital Information*, 1(5), January 2000.

[11] C. Monz, J. Kamps, and M. de Rijke, 'The University of Amsterdam at CLEF 2002', *Working Notes of the CLEF 2002 Workshop*, pp. 73-84, 2002.

[12] 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.

[13] M. Porter, 'Snowball: A Language for Stemming Algorithms', http://snowball.tartarus.org/texts/introduction.html, (visited 13 March 2003).

[14] D. Reidsma, D. Hiemstra, F. de Jong, and W. Kraaij, 'Cross-language Retrieval at Twente and TNO', *Working Notes of the CLEF 2002 Workshop*, pp. 111-114, 2002.

[15] J. Savoy Cross-language information retrieval: experiments based on CLEF 2000 corpora. *Information Processing and Management*, 39(1):75-115, 2003.

[16] S. Tomlinson, 'Experiments in 8 European Languages with Hummingbird SearchServer at CLEF 2002', *Working Notes of the CLEF 2002 Workshop*, pp. 203, 214, 2002.

[17] C. Zhai and J. Lafferty, 'A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval' *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 334-342, 2001.

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