Exploring New Languages with HAIRCUT at CLEF 2005

Paul McNamee
Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, MD 20723-6099 USA
paul.mcnamee@jhuapl.edu

JHU/APL has long espoused the use of language-neutral methods for cross-language information retrieval. This year we participated in the ad hoc cross-language track and submitted both monolingual and bilingual runs. We undertook our first investigations in the Bulgarian and Hungarian languages. In our bilingual experiments we used several non-traditional CLEF query languages such as Greek, Hungarian, and Indonesian, in addition to several western European languages. We found that character n-grams remain an attractive option for representing documents and queries in these new languages. In our monolingual tests n-grams were more effective than unnormalized words for retrieval in Bulgarian (+30%) and Hungarian (+63%). Our bilingual runs made use of *subword translation*, statistical translation of character n-grams using aligned corpora, when parallel data were available, and web-based machine translation, when no suitable data could be found.

Subject Descriptors

H.3.1 **[Information Systems]**: Content Analysis and Indexing – linguistic processing, indexing; H.3.3 **[Information Systems]** : Information Search and Retrieval – query formulation.

Keywords

Cross-language information retrieval, character n-gram tokenization, pre-translation query expansion, parallel corpora, translation.

Introduction

HAIRCUT[1] is a Java-based information retrieval system that has been developed at the Johns Hopkins University Applied Physics Laboratory. An early version of HAIRCUT was created for use in the TREC-6 evaluation. One of the original issues that we wanted to investigate with the HAIRCUT system was whether character n-gram tokenization was an effective technique for ad hoc text retrieval. Earlier work using n-grams had been viewed with skepticism [3] and it was our intent to compare n-grams and words in an identical framework (*i.e.,* keeping the retrieval system constant). Our early results were promising and we found that the use of n-grams conveys substantial advantages when non-English collections were used [7].

JHU/APL was a participant in the first CLEF evaluation, and since then, we have been able to apply our techniques in the ten languages explored in the ad hoc tasks, as well as in Chinese, Japanese, and Korean (at NTCIR), and Arabic (at TREC). We have found n-gram tokenization to be surprisingly effective across these diverse languages. We believe n-grams are effective, in part, because they account for morphological variation and provide robustness in the face of slight orthographic mismatching. N-grams also obviate the need to perform decompounding (*e.g.,* in German) or word segmentation (*e.g.,* in Chinese).

In addition to the use of character n-gram tokenization we make use of a statistical language model of retrieval and combination of evidence from multiple retrievals. For bilingual retrieval we include pre-translation query expansion using comparable collections, statistical translation from aligned parallel collections, and when translation resources are scarce, reliance on language similarity alone. This year we continue experimenting with a technique we first applied at the CLEF 2003 evaluation: *subword translation*, translation of the constituent n-grams in queries rather than words [8]. For translation we used aligned parallel corpora instead of bilingual wordlists, when possible, and other resources (*e.g.,* Web-based MT) when not. Subword translation attempts to overcome obstacles in dictionary-based translation, such as word

[1] HAIRCUT stands for the Hopkins Automated Information Retriever for Combing Unstructured Text.

lemmatization, matching of multiple word expressions, and inability to handle out-of-vocabulary words such as common surnames [12].

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 sometimes be obtained.

Methods

HAIRCUT supports several ways of representing documents using an order independent, bag-of-terms model. Note we are frequently using 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 using the Snowball stemmer). 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 efficiency, and because such terms typically add little to query performance. (By default, query terms occurring in greater than 20% of documents are ignored).

We continue to use a statistical language model for retrieval akin to those presented by Ponte and Croft [13] and Hiemstra [4] with Jelinek-Mercer smoothing [5] (*i.e.,* linear interpolation). In this model, the probability of relevance is given as:

$$P(D \mid Q) = \underset{q\,?Q}{?}\left[aP(q \mid D) + (1-a)P(q \mid C)\right],$$

where Q is a query, D is a document, C is the collection as a whole, and a 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. It has been 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.

Character n-grams, sequences of *n* consecutive characters, have been used for a number of tasks in human language technology (*e.g.,* spelling correction [14], diacritics restoration [11], and language identification [1]). Their use for IR dates to the mid-1970s where they were used primarily as a technique to decrease dictionary size. At that time *n=2* or *n=3* were typical lengths, and for a fixed alphabet size a substantial reduction in memory requirements could be realized. Over time as physical memory costs fell significantly, research in the mid-1990s led to n-grams being considered as an alternative indexing representation to words or stemmed words (see [3]). There are several variations on n-gram indexing; here we concentrate on overlapping character n-grams of a fixed length (typically *n=4* or *n=5*). For the text ‘prime_minister’ and *n=7* the resulting n-grams are: ‘_prime_’, ‘prime_m’, ‘rime_mi’, ‘ime_min’, ‘me_mini’, ‘e_minis’, ‘minist’, ‘ministe’, ‘inister’, and ‘nister_’. The single n-gram ‘ime_min’ that occurs at the word boundary is fairly distinct indicator of the query phrase ‘prime minister’ and it would not be generated from a sentence like ‘the finance minister ordered prime rib for lunch’ which might cause a false match using words alone as indexing terms.

Monolingual Task

For our monolingual work we created 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 (when available), 4-grams, and 5-grams. Information about each index is shown in Table 1 (below).

Selection of 4-grams and 5-grams as indexing terms was based on a comprehensive study across the CLEF languages that investigated n-gram length [9] and established that 4-grams and 5-grams seem to work equally well for monolingual retrieval. Our language model requires a single smoothing constant; we used ? =0.3 with both words and stems, and ? =0.5 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 from the top 1000). Figure 1 charts performance using our four different term indexing strategies, in isolation. In the Bulgarian and Hungarian languages, substantial benefits were seen when n-grams were used – 30% and 63% relative improvements, respectively. In the other languages, n-grams performed similarly to words and somewhat worse than the use of stemmed words (*e.g.*, in English and French). Our previous experience has shown that n-grams produce larger benefits in languages with greater morphological complexity.

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
BG	67341	778	57 / 67	---	154 / 193	251 / 769
EN	166754	2063	143 / 302	123 / 236	504 / 166	827 / 916
FR	177450	2537	129 / 328	107 / 226	393 / 159	628 / 838
HU	49530	939	59 / 549	---	121 / 150	200 / 741
PT	210734	2904	178 / 418	140 / 254	529 / 174	868 / 907

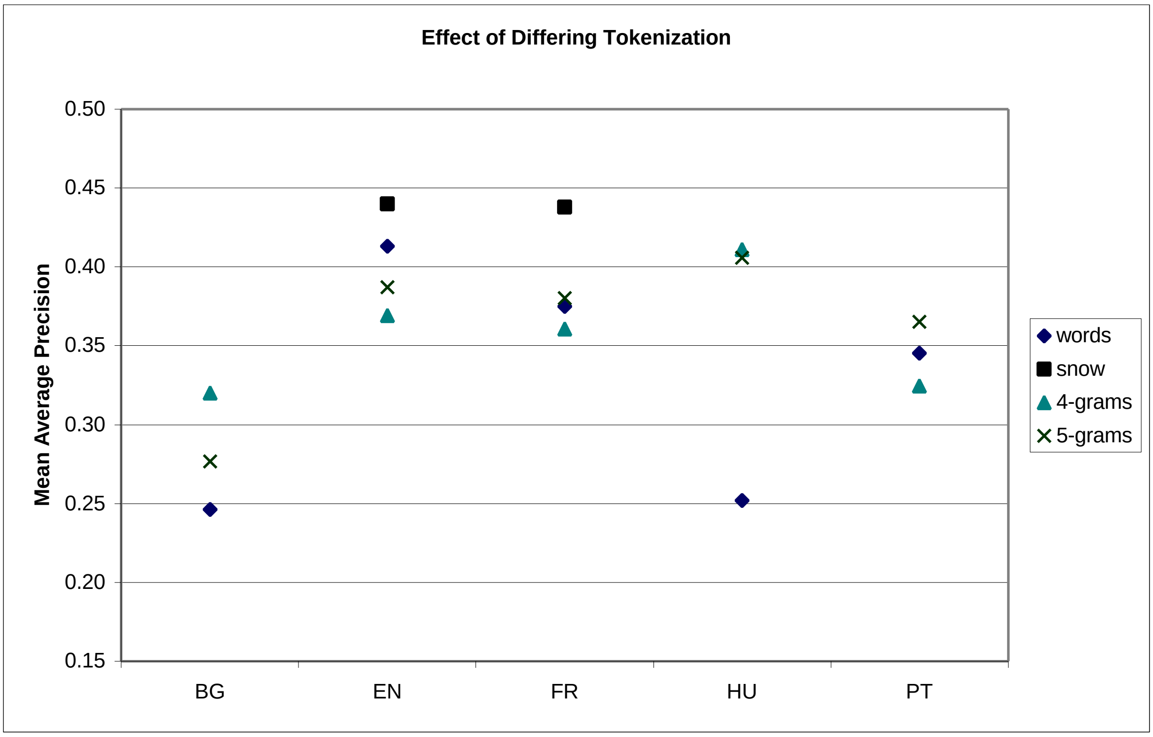

Figure 1. Relative effectiveness of tokenization methods on the CLEF 2005 test sets.

Our submitted runs were based on a combination of several base runs using various options for tokenization. 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. We produced three to five runs in each language that were created from combinations of the base runs. Runs were labeled *aplmoxx[a-e]*, where *xx* indicates the language of interest. Runs whose names end with a terminal 'a' were produced by combining a 5-gram base run with a stemmed word base run; a terminal 'b' indicates fusion of a 4-grams and stemmed words; terminal 'c' is used for runs that used both 4-grams and 5-grams; the suffix 'd' indicates solitary use of 4-grams; and, a terminal 'e' indicates the use of 5-grams alone. Monolingual performance based on mean average precision is reported in Table 2.

Table 2. Official results for monolingual task.

run id	Fields	Terms	MAP	Rel. Found	Relevant
aplmobgc	TD	4+5	0.3058	706	778
aplmobgd	TD	4	0.3203	678	778
aplmobge	TD	5	0.2768	699	778
aplmoena	TD	5+snow	0.4346	1930	2063
aplmoenb	TD	4+snow	0.4222	1900	2063
aplmoenc	TD	4+5	0.3898	1877	2063
aplmoend	TD	4	0.3692	1808	2063
aplmoene	TD	5	0.3873	1889	2063
aplmofra	TD	5+snow	0.4114	2422	2537
aplmofrb	TD	4+snow	0.4122	2427	2537
aplmofrc	TD	4+5	0.3765	2283	2537
aplmofrd	TD	4	0.3608	2109	2537
aplmofre	TD	5	0.3801	2274	2537
aplmohuc	TD	4+5	0.4063	893	939
aplmohud	TD	4	0.4112	893	939
aplmohue	TD	5	0.4056	891	939
aplmoptc	TD	4+5	0.3610	2446	2904
aplmoptd	TD	4	0.3246	2343	2904
aplmopte	TD	5	0.3654	2450	2904

Bilingual Task

Our preferred approach to bilingual retrieval is based on the following procedure: (1) apply pre-translation query expansion using the source language CLEF corpus; (2) translate terms statistically using aligned parallel corpora, where terms can be words, stems, or n-grams; (3) and, perform retrieval using the query terms that were projected into the target language, possibly with additional relevance feedback. We have had good success using aligned parallel corpora to extract statistical translations. Others have also relied on corpus-based translation; however, we recently demonstrated significant improvements in bilingual performance by translating character n-grams directly. We call this '*subword translation*'. Additionally we also translate stemmed words and words. This year we were only able to use this technique for the English, French, and Portuguese target collections as we lacked parallel resources in Bulgarian and Hungarian.

For the 2002 and 2003 campaigns we relied on a single source for parallel texts, the Official Journal of the E.U. [15], 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 most of 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. We managed to use this data for stem-to-stem translation in the CLEF 2004 evaluation and we used this data again this year for word, stem, and n-gram translation.

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 Church's *char_align* [2].

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 (*e.g.*, symmetric conditional probability or mutual information) 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 retrieval and translation.

We performed pairwise alignments between languages pairs, for example, between English and Portuguese. Once aligned, we indexed each pairwise-aligned collection using the technique described earlier on the CLEF-2005 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. So far we have been using 1-best translation, but we can generate multiple weighted translations for each term. 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. Our experience on the CLEF 2003 and 2004 bilingual test sets 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 might yield an improvement [10].

Unfortunately, our data from the Official Journal of the EU did not cover two of the target language collections (*i.e.*, Bulgarian and Hungarian). To support translation to or from these languages we relied on query translation using web-based machine translation. We also used MT to use the Greek and Indonesia query sets against English documents. The online services we used are located at:

- http://babelfish.altavista.com (GR to EN)
- http://www.toggletext.com/kataku_trial.php (IN to EN)
- http://www.bultra.com/test_e.htm (BG to/from EN)
- http://www.tranexp.com/ (HU to/from EN)

As can be seen in Table 3 (below), our results using corpus-based subword translation achieved bilingual performance between 78% and 87% of our best monolingual runs for the given target language. Table 4 details our results using available machine translation software. The resultant bilingual performance depends heavily on the individual translation engine used (from 26% to 85% of our best monolingual baselines). In some cases the result of fusing multiple runs using different target-side tokenization of the machine translation output resulted in an improvement, for example, run *aplbiidend* had a 4% absolute improvement in mean average precision of *aplbiidena*, which used 5-grams alone. In a couple of cases we directly compared the use of 4-grams and 5-grams on the MT output and found the results to be very similar (e.g., compare *aplbienbg[a/e]* and *aplbienhu[a/e]*).

Table 3. JHU/APL's official results for bilingual task using corpus-based translation.

run id	Source	Target	Fields	Terms	MAP	% mono	Rel. Found	Relevant
aplbienfrc	EN	FR	TD	5-grams	0.3442	78.62%	2108	2537
aplbienptb	EN	PT	TD	5-grams	0.3130	85.39%	2053	2904
aplbiesptb	ES	PT	TD	5-grams	0.3185	87.16%	2268	2904

Table 4. JHU/APL's official results for bilingual task using machine translation.

run id	Source	Target	Fields	Terms	MAP	% mono	Rel. Found	Relevant
aplbigrena	GR	EN	TD	5-grams	0.2418	54.94%	1388	2063
aplbihuena	HU	EN	TD	5-grams	0.1944	44.17%	1363	2063

aplbiidena	ID	EN	TD	5-grams	0.3313	75.28%	1698	2063
aplbiidend	ID	EN	TD	w/s/4/5	0.3728	84.71%	1796	2063
aplbienbga	EN	BG	TD	5-grams	0.0833	26.01%	438	778
aplbienbge	EN	BG	TD	4-grams	0.0959	29.94%	423	778
aplbienhua	EN	HU	TD	5-grams	0.2235	54.35%	718	939
aplbienhue	EN	HU	TD	4-grams	0.2458	59.78%	729	939

In Bulgarian and Hungarian it seems that 4-grams may have a slight advantage over 5-grams, though additional testing should be performed to verify that the differences are statistically significant. However, the use of n-grams over raw words seems clearly indicated.

Conclusions

JHU/APL participated in the ad hoc tasks in the CLEF 2005 evaluation, using our language-neutral approach that prominently features character n-gram tokenization and statistical translation using aligned parallel corpora. This year we had to rely on web-based machine translation for mappings between several language pairs, for which we had been unable to obtain suitable parallel data. We compared words, a popular suffix stemmer, and n-grams of lengths four and five on the monolingual collections, all using the same retrieval engine and language model similarity metric. We found that n-grams continued to work well for monolingual retrieval, though their superiority was only apparent in Bulgarian and Hungarian.

We continued to combine runs produced through disparate retrievals, which, in the past, we have seen a modest (*e.g.*, 10% relative) improvement. This year, however, we noted that our single-best tokenization method outperformed merging of disparate runs (compare Figure 1 and the results in Table 2).

For bilingual retrieval we employed subword translation in several official runs, with good effect. However we still lack parallel corpora for Bulgarian and Hungarian. We would like to expand on these experiments if we can locate appropriate data. Our results from this year agree with previous findings that character n-grams remain effective and an attractive alternative, especially in languages with complex morphology or ones in which resources (*e.g.*, morphological analyzers or stemmers) are difficult to obtain or use. Our recipe for bilingual retrieval appears effective, but is best accomplished when parallel data are available.

References

[1] W. B. Cavnar and J. M. Trenkle, 'N-Gram Based Text Categorization.' In: *Proceedings of the Third Symposium on Document Analysis and Information Retrieval*, pp. 161-169, 1994.

[2] 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.

[3] M. Damashek, 'Gauging Similarity with n-grams: Language-Independent Categorization of Text.' *Science*, 267:843-848, 1995.

[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] J. Mayfield, P. McNamee and C. Piatko, "The JHU/APL HAIRCUT System at TREC-8". In E. Voorhees and D. Harman (eds.), *Proceedings of the Eighth Text REtrieval Conference (TREC-8*), NIST Special Publication 500-246, Gaithersburg, Maryland, 2000.

[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] P. McNamee and J. Mayfield, 'Translating Pieces of Words.' *Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval (SIGIR-2005)*, Salvador, Brazil, pp. 643-644, August 2005.

[11] R. Mihalcea and V. Nastase, 'Letter Level Learning for Language Independent Diacritics Restoration.' In: *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pp. 105-111, 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] J. M. Ponte and W. B. Croft, ‘A Language Modeling Approach to Information Retrieval.’ In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 275-281, 1998.

[14] E. M. Zamora, J. J. Pollock, and A. Zamora, ‘The Use of Trigram Analysis for Spelling Error Detection.’ *Information Processing and Management 17:305-316, 1981.*

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