Indexing Using Both N-Grams and Words

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

Goals

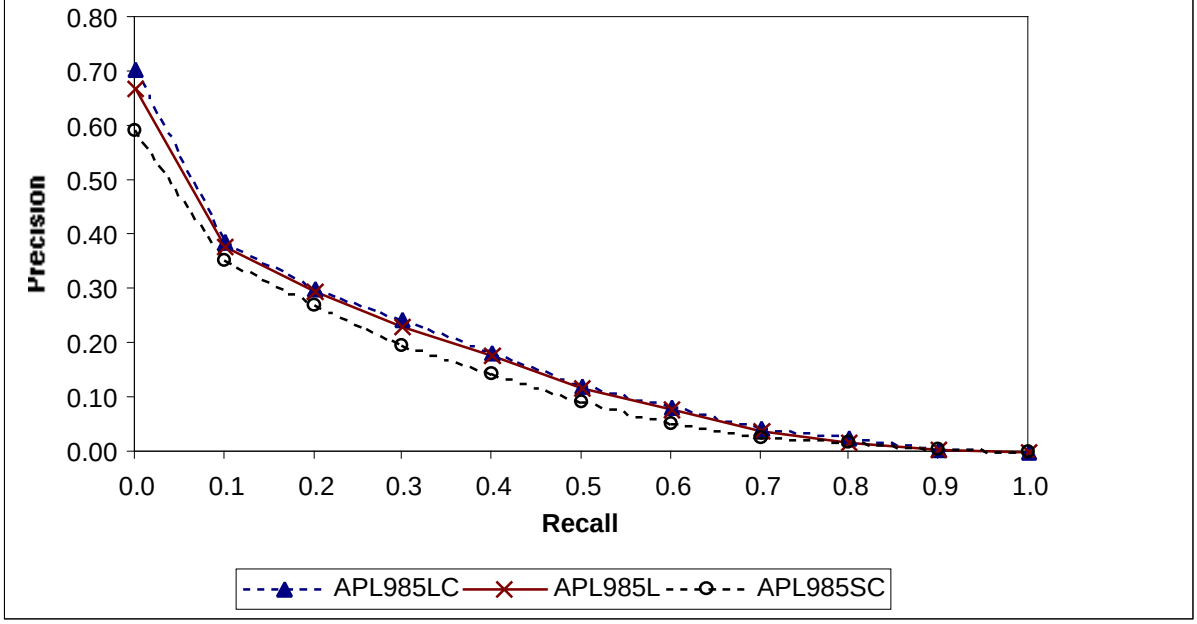
The Johns Hopkins University Applied Physics Laboratory (JHU/APL) is a first-time entrant in the TREC Category A evaluation. The focus of our information retrieval research is on the relative value of and interaction among multiple term types. In particular, we are interested in examining both words and n-grams as indexing terms. The relative values of words and n-grams have been disputed; to our knowledge though, no one has previously studied their relative merits while holding all other aspects of the system constant.

Approach

The Hopkins Automated Information Retriever for Combing Unstructured Text (HAIRCUT) system was built to explore the use of multiple term types. The system is implemented in Java for ease of development, portability, and of course blazing speed. It implements a vector model, using cosine as its similarity measure. Terms are usually weighted by Okapi BM 25 [Walker *et al.*, 1998], which is a variant of TF/IDF weighting that boosts the scores of longer documents. Normal TF/IDF and plain TF weightings are also supported. Cosines can be computed either relative to the origin, or relative to the corpus centroid. Terms that appear in a high percentage of documents are effectively stop-listed.

HAIRCUT performs rudimentary preprocessing on queries to remove stop structure [Allan *et al.*, 1998], *e.g.*, affixes such as "… would be relevant" or "relevant documents should…." Other than this preprocessing, queries are parsed in the same fashion as are documents in the collection.

Figure 1. TREC-7 Ad hoc task, APL Official Runs



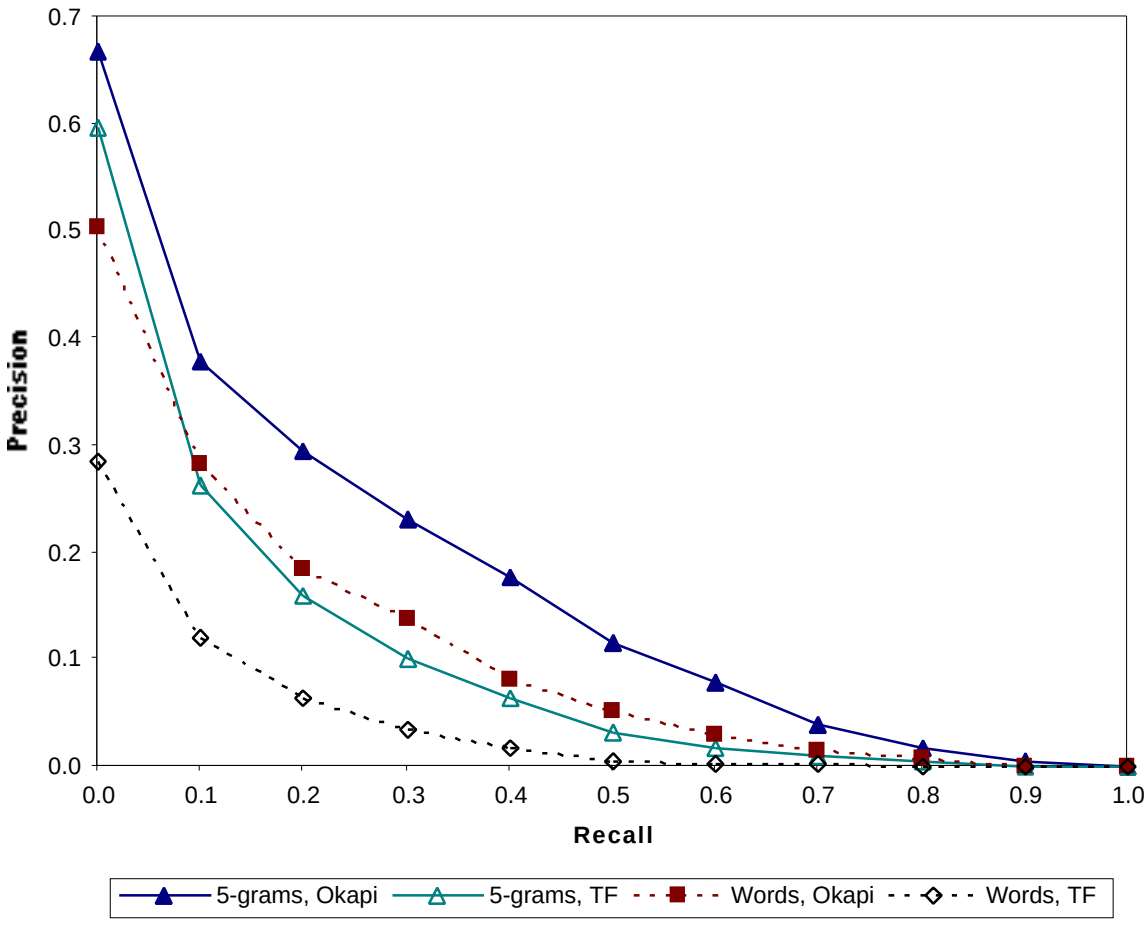
We conducted our work on a pair of Sun workstations, an UltraSPARC 2 with two 300 MHz processors and a 4-node Ultra Enterprise 450 server. Both workstations had 512 MB of physical memory and access to 26 GB of shared hard disk space.

The HAIRCUT system comprises approximately 28,000 lines of Java code.

For TREC-7 we tested two types of terms: words and 5-grams. After eliminating punctuation, downcasing letters, and mapping numbers to a single digit, a word was any remaining blank-delimited sequence of characters.

For n-grams we used 5-grams formed from the same character stream used for selecting words, but with common words replaced by a single character. Although Java uses the 2-byte Unicode format to represent strings, HAIRCUT represents terms using byte sequences. Since the input stream is downcased, all uppercase letters and certain of the Latin-1 characters can be used as replacements for common words such as "the," "with," *etc*. This has an effect of lengthening n-grams that span common words. For example, the phrase "statue_of_liberty" might produce the 5-gram "e_¢_l" in HAIRCUT, where the common word 'of' has been replaced by the single character '¢'.

Figure 2. TREC-7 Ad hoc task, Long Topics, Words *vs.* 5-grams



Ad hoc Results

JHU/APL submitted three ad hoc runs. Our first run, labeled APL985L, was a baseline run that used 5-grams as indexing terms, and used no relevance feedback. Runs APL985LC and APL985SC combine two separate query runs; a 5-gram run and a word-based run that used automated relevance feedback. APL985L and APL985LC used the title, description, and narrative portions of the topic statements, while APL985SC only made use of the title section.

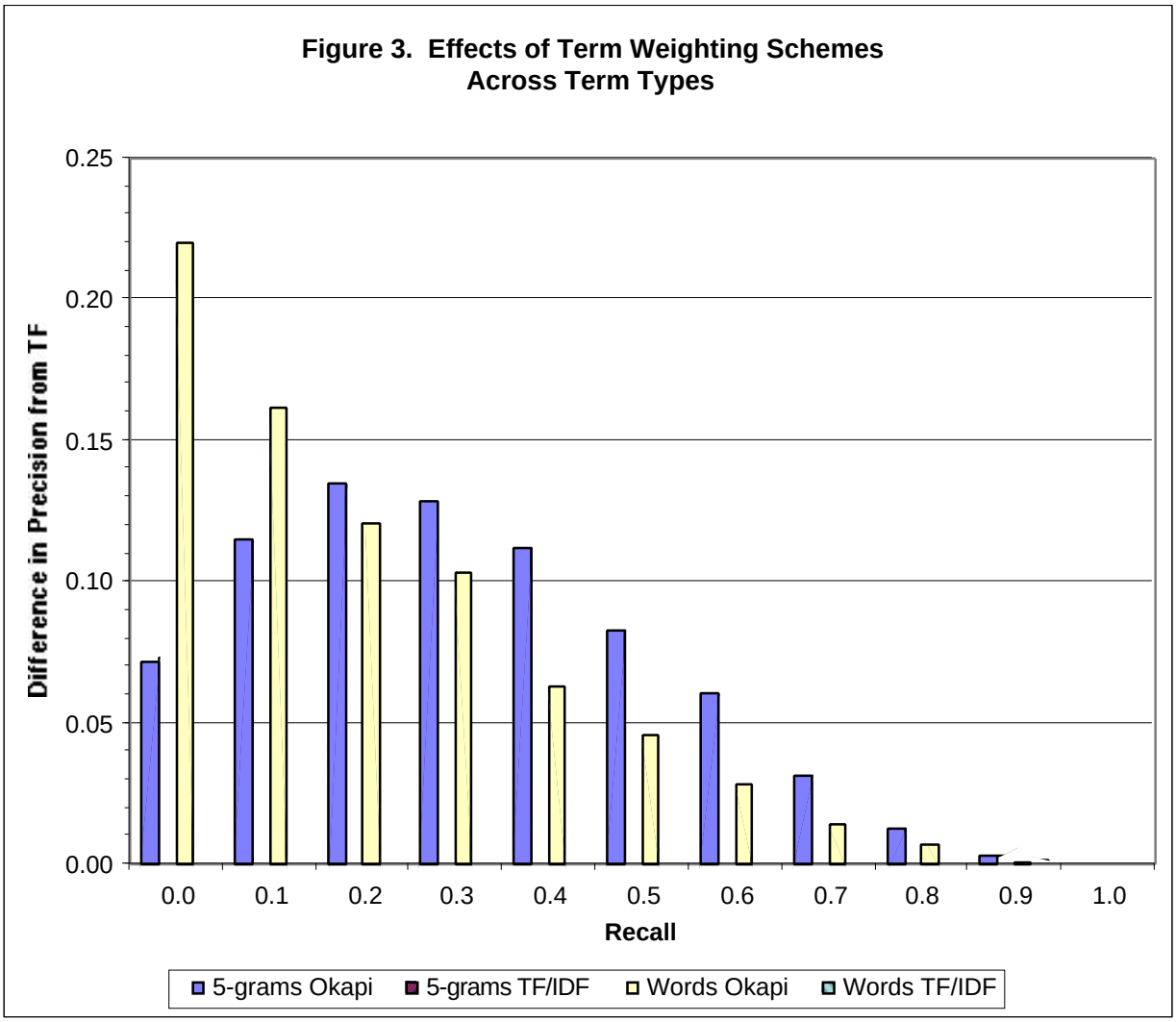
This year we concentrated our efforts on techniques that improve precision at low recall levels. Our official results are shown in Figure 1. The APL985L and APL985LC runs are similar, showing that our relevance feedback techniques were ineffectual on queries composed of the title, descriptive, and narrative portions of the topic statements.

N-Grams vs. Words

One of our main goals in developing HAIRCUT has been to compare the relative merits of n-grams and words as indexing terms when all other aspects of the system are held constant. To that end, we scored a set of word-based runs against the TREC-7 relevance assessments.

Figure 2 depicts the results of our experiments comparing the use of unstemmed words against 5-grams. Surprisingly good performance was obtained from the 5-grams. In fact, 5-grams using term frequency weighting do about as well as words using Okapi BM 25 term weighting. None of these runs uses relevance feedback.

Figure 3. Effects of Term Weighting Schemes Across Term Types



Other Experiments

In addition to comparing n-grams and words as term types, we also wanted to investigate the effects of alternative term weighting schemes and the effects of using cosine with corpus centroid subtraction as a measure of similarity. We performed a series of runs on the TREC-7 topics to explore the variability in performance attributable to these alternative approaches.

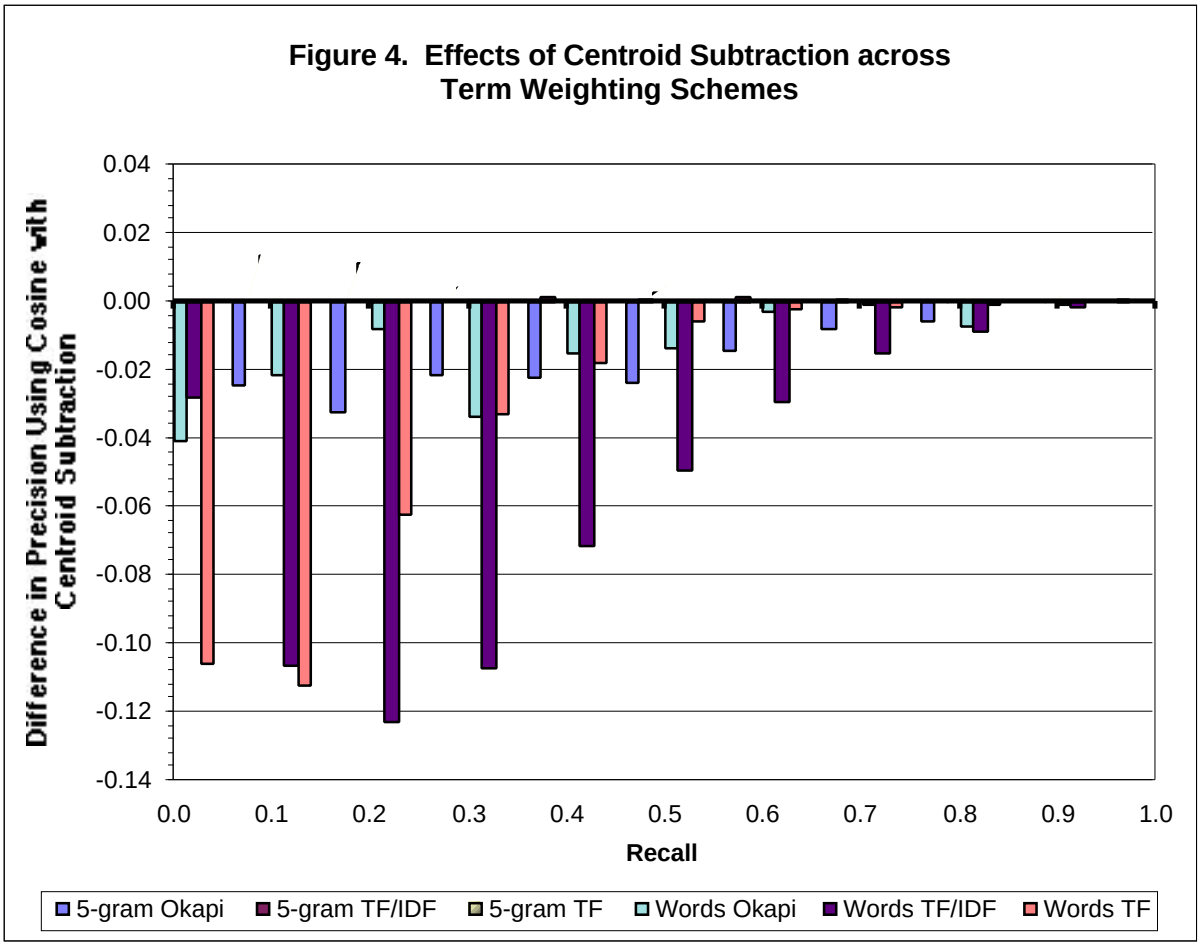
As expected, both Okapi-style term weighting and TF/IDF term weighting surpass the use of term frequencies alone. Figure 3 illustrates the gain from the use of Okapi and TF/IDF. Figure 4 shows the effects of adding centroid subtraction to term weighting.

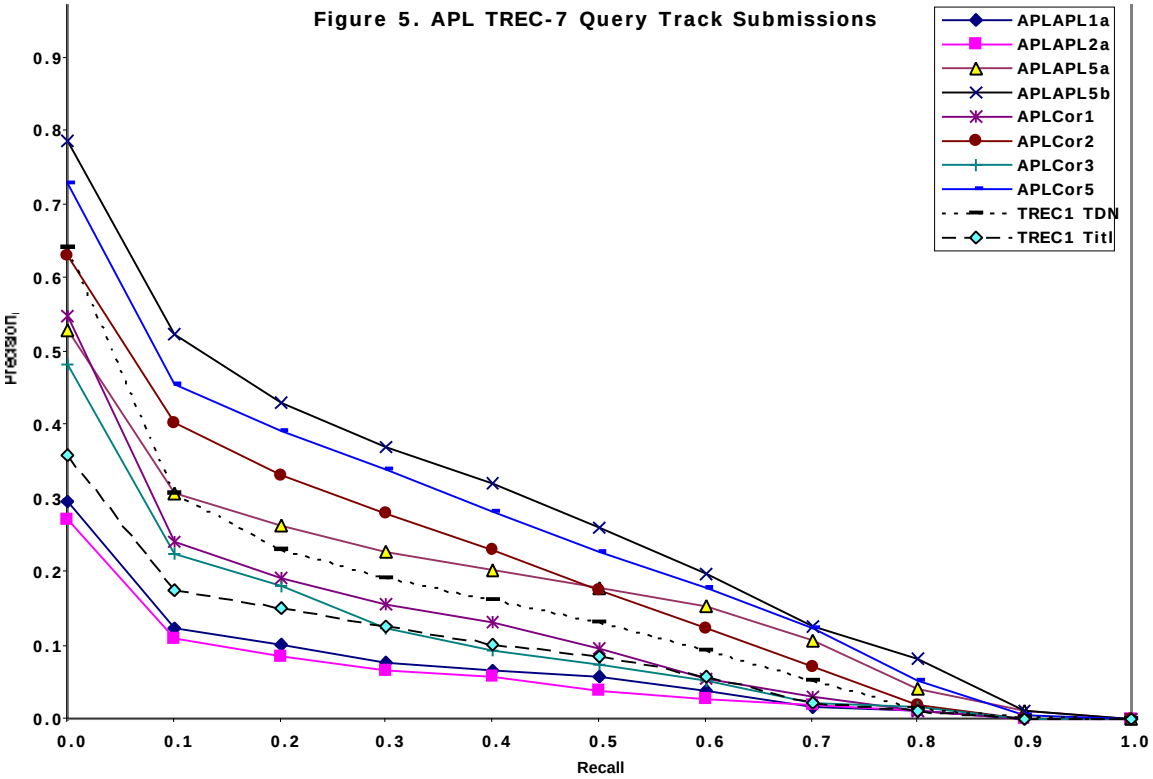
Four observations are worth making about these data. First, Okapi and TF/IDF term weighting are very nearly equal in terms of their improvement from TF weighting. Secondly, the use of Okapi and TF/IDF appears to provide more of a gain for word-based runs than for n-gram runs at lower recall levels. Correspondingly, the n-gram based runs appear to receive more advantage from the alternative term weightings at higher recall levels. Thirdly, these results are based on unofficial runs, for which the top one hundred documents are not guaranteed to have been assessed. The average number of documents in the top one hundred that were assessed in the Okapi and TF/IDF runs ranges from 92% to 95%, but the TF runs averaged only 78% for n-grams and 66% for words. It is conceivable that some of the difference reported here stems from differences in level of assessment. Finally, we have yet to see circumstances under which the use of centroid subtraction is justified.

Query Track

JHU/APL also participated in the query track. We generated four query sets. Two (APL1a and APL2a) were generated by hand, by reading the narrative version of each source query and generating a title query and a description query for each. The third (APL5a) was created using a

Figure 4. Effects of Centroid Subtraction across Term Weighting Schemes





variant of the mutual information statistic to extract important terms from the top 75 documents retrieved for the source query. The last (APL5b) used the same statistic to extract important terms from the query track training set. All terms in the last two query sets were unstemmed words; we did not anticipate that other systems could make use of n-grams.

Our goal this year was simply to assess the variability in precision found across queries. To that end, we used a single system configuration to process eight of the nine query track query sets (one query set from Cornell included a Boolean component, which our system cannot handle). This configuration used unstemmed words as terms, and cosine based at the origin to gauge document similarity. No relevance feedback was used.

Our results, shown in Figure 5, exhibited tremendous variability in result quality across the eight query sets. The best results were obtained from the two query sets developed using training data. The query sets that we generated by hand after reading the source narratives fared worst. Figure 5 also shows our results on the original TREC-1 queries, both title-only and title-description-narrative. We are currently trying to assess the relative contributions of vocabulary choice, lack of assessments, and system configuration to this range of results.

References

[Allan *et al.*, 1998] James Allan, Jamie Callan, W. Bruce Croft, Lisa Ballesteros, Don Byrd, Russell Swan, and Jinxi Xu, 'INQUERY does battle with TREC-6.' In E. M. Voorhees and D. K. Harman, eds., *The Sixth Text REtrieval Conference (TREC-6).* NIST Special Publication 500-240, pp. 169-206, 1998.

[Walker *et al.*, 1998] S. Walker, S. E. Robertson, M. Boughanem, G. J. F. Jones and Karen Sparck Jones, 'Okapi at TREC-6, Automatic ad hoc, VLC, routing, filtering and QSDR.' In E. M. Voorhees and D. K. Harman, eds., *The Sixth Text REtrieval Conference (TREC-6).* NIST Special Publication 500-240, pp. 125-136, 1998.