N-Grams vs. Words as Indexing Terms

James Mayfield
Paul McNamee
The Johns Hopkins University
Applied Physics Laboratory
Johns Hopkins Rd.
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 evaluation. We entered Category B, which restricted indexing and retrieval to the Financial Times corpus.

We had two goals for our TREC entry: first, to build a working information retrieval system in Java capable of supporting a variety of term types and weighting methods; and second, to use the system to evaluate the relative merits of words, stemmed words, and n-grams as terms for information retrieval. The relative merits of words and n-grams as retrieval terms have been disputed; to our knowledge though, no one has 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 address these goals. The system consists of roughly 7000 lines of Java code. It implements a vector model, using cosine as its similarity measure. Terms that appear in more than 2% of documents are effectively stop-listed.

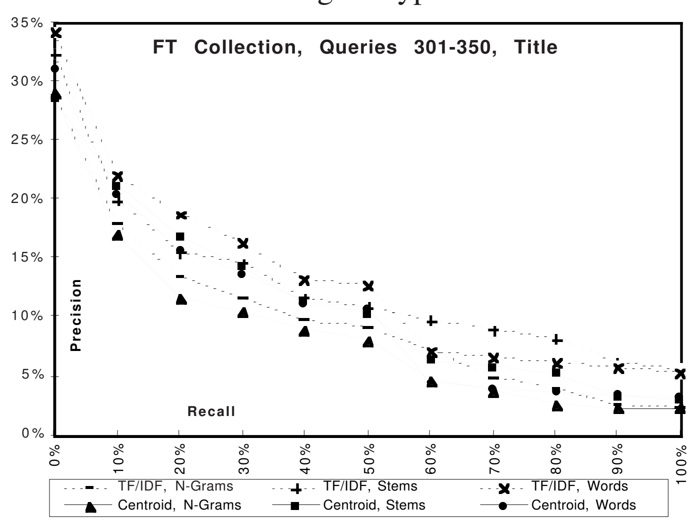
We tested three types of terms: words, stemmed words, and n-grams. After eliminating punctuation, downcasing letters, and mapping numbers to a single digit, a word was any remaining blank-delimited sequence of characters. Stems were produced by feeding this word stream to a stemming algorithm; our stemmer produced results virtually identical to those of the SMART version 11 stemmer. For n-grams we used 5-grams, formed from the same character stream used for selecting words.

We also tested two term-weighting methods. Term frequency/inverse document frequency (TF/IDF) is a standard weighting scheme in word-based systems; we used the term frequency within a document times the log of the total number of documents over the number of documents containing the term. Several n-gram-based systems have used a competing term-weighting scheme, in which the centroid of the entire collection is subtracted from the vector representing a document. This transformation effectively moves the base of the angle used in the cosine calculation from the origin to the corpus centroid. The HAIRCUT system allowed us to compare these two weighting methods while holding all other aspects of the system unchanged.

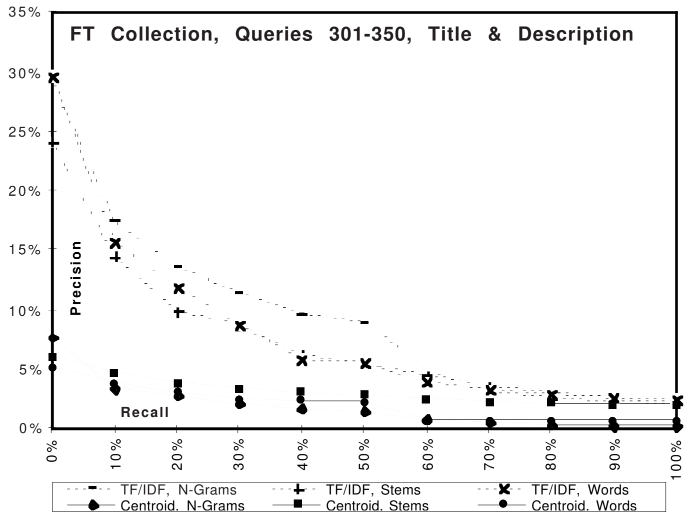
Results

We ran each of the six term-type/weighting method pairs against otherwise unprocessed queries of three lengths: short queries comprising just the title words, medium-length queries comprising both title and description, and long queries that included the entire narrative (but not the title or description). The results varied significantly with the length of the query.

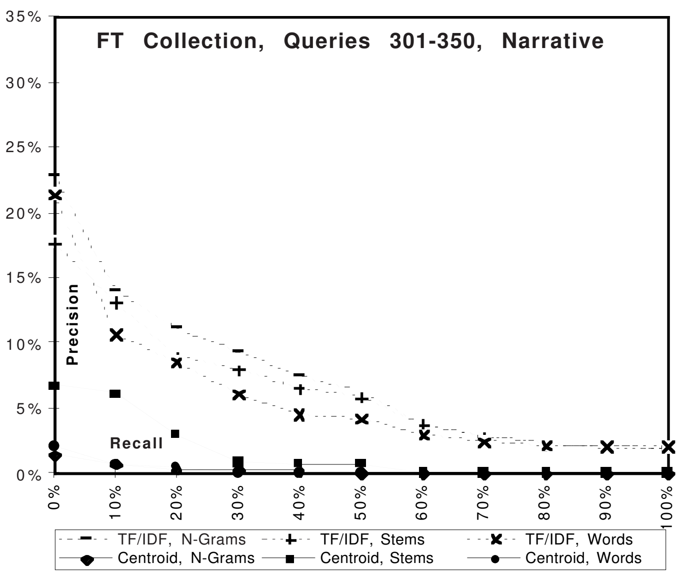
Short queries led to roughly similar performance from the six configurations. However, two trends emerged. First, words and stemmed words proved superior to n-grams. Second, TF/IDF exhibits a small but almost universal advantage over centroid subtraction when holding the type of term fixed:



In moving to medium-length queries, one of these trends is amplified, while the other is reversed. The superiority of TF/IDF over centroid subtraction is now quite pronounced. On the other hand, n-grams now produce better performance than words or stemmed words. Note that overall performance is lower for medium-length queries than for short queries:



The trends seen in the above graph remain unchanged when long queries are used. The gulf between TF/IDF and centroid subtraction remains wide, while N-grams retain their slight edge over words:



Conclusions

We draw two main conclusions from our experiments:

1. In a simple cosine-based system, words or stemmed words are better for short queries, while n-grams are better for long queries.
2. TF/IDF term weighting is uniformly superior to centroid subtraction term weighting.

These conclusions of course do not argue definitively for using a particular type of term, or even for a particular form of term weighting. After all, each of these approaches has advantages beyond its contribution to improved precision/recall. For example, words lend themselves to use with semantics-based techniques that n-grams are unlikely to support. Furthermore, indexes built using n-grams are significantly larger than those built using words or stems. On the other hand, n-grams exhibit language-independence and error-tolerance that can make them the terms of choice for some applications. Given that words and n-grams are not too far apart on the recall/precision graphs, the choice of term type is probably better made based on other considerations.

The situation with term weighting is clearer; TF/IDF is the clear winner over centroid subtraction. Even here, though, the usefulness of a corpus centroid in other areas (*e.g.,* as metadata about the corpus as a whole) might prove to be an overriding factor.

And of course, all of these conclusions must be viewed in the light of the relatively low performance of the HAIRCUT system overall. That performance was to be expected; the system, which was written from scratch, became operational within a few scant hours of the TREC-6 submission deadline (in fact, 24 hours earlier it exhibited worse-than-random performance). However, we believe that the test results are indicative of important trends. We will rerun these tests as the HAIRCUT system matures.