Strength and Similarity of Affix Removal Stemming Algorithms

William B. Frakes
Computer Science Department
Virginia Tech

Christopher J. Fox
Computer Science Department
James Madison University

Abstract

This study evaluated the strength of, and similarity among, four affix removal stemming algorithms. Strength and similarity were evaluated in different ways, including new metrics based on the Hamming distance measure. Data was collected on stemmer outputs for a list of 49,656 English words derived from the UNIX spelling dictionary and the Moby corpus. Conclusions about the relative strength and similarity of the four stemming algorithms are reported.

Introduction

Stemming algorithms, or stemmers, are used to group words based on semantic similarity. There are several types of stemming algorithms. Affix removal algorithms are the most common. Affix removal stemming algorithms remove affixes (suffixes or prefixes) from words producing a root form called a stem that often closely approximates the root morpheme of a word.

Stemming algorithms are used in many types of language processing and text analysis systems, and are also widely used in information retrieval and database search systems. There have been many studies of conflation for information retrieval systems as summarized, for example, in [Frakes, 92]. Most of these studies have focused on the effect of stemming on retrieval performance measured with recall and precision. A few studies have also looked at stemming as a method for index compression.

The closeness of a stem to its corresponding root morpheme is sometimes used as a measure of stemmer correctness. Another measure of stemmer correctness is how many semantically related words are correctly assigned to a conflation class. While stemmer correctness is important, it is not the focus of this study. Rather, we address the problem of how stemming algorithms differ, so that one can choose among them for a particular application. We address this issue by evaluating the similarity of four affix removal stemming algorithms using a variant of the Hamming distance function and other measures.

Stemmer Strength Metrics

The degree to which a stemmer changes words that it stems is called stemmer strength. For affix removal stemmers, such a change is either the removal of characters corresponding to an affix, for example reducing the word *engineering* to the stem *engineer*, or recoding. Recoding is replacing one letter with another, for example changing y to I so that the word *sky* becomes the stem *ski*.

Stemmer strength is important because, especially in the character removal case, it can be predictive of recall and precision and of index compression. A stronger stemmer will, on average, increase recall, decrease precision, and increase index compression. Some ways to measure stemmer strength are:

- The mean number of words per conflation class—This is the average number of words that correspond to the same stem for a corpus. For example if the words "engineer," "engineered," and "engineering" are stemmed to "engineer," then this conflation class size is three. Stronger stemmers will tend to have more words per conflation class.
- Index compression factor—The index compression factor is defined as (n-s)/n where n is the number of words in the corpus and s is the number of stems. In other words, the index compression factor is the fractional reduction in index size achieved through stemming. For example, a corpus with 50,000 words (n) and 40,000 stems (s), would have an index compression factor of 20%. Stronger stemmers will tend to have larger index compression factors.
- The number of words and stems that differ—Stemmers often leave words unchanged. For example, a stemmer might not alter "engineer" because it is already a root word. Stronger stemmers will change words more often than weaker stemmers.
- The mean number of characters removed in forming stems—Stronger stemmers remove more characters from words to form stems. For example, a stemmer that stems the corpus {engineer, engineered, engineering, engineers} to the stem engineer would remove an average of (0+2+3+1)/4 = 1.5 characters. A weakness of this metric is that it does not measure transformations of stem endings. We therefore have developed the following measures.
- The median and mean modified Hamming

distance between words and their stems—The Hamming distance between two strings of equal length is defined as the number of characters in the two strings that are different at the same position. For strings of unequal length we add the difference in length to the Hamming distance to give a modified Hamming distance function d. This measure takes into account transformations of stem endings. For example, a stemming algorithm might reduce the corpus { try, tried, trying } to the stem tri. The mean modified Hamming distance between the original words and the stem is (1+2+4)/3 = 2.33 characters, and the median is 2.

These six metrics are used to compare the strength of the four stemmers.

Using these stemmer strength measures, it is possible to define the limits of stemmer strength. The strongest possible affix removal stemmer would be one that removed all but the first character from each stemmed word. The number of conflation classes in this case would be 26, the mean conflation class size would be the number of words in the corpus divided by 26, and the compression factor would be (n-26)/n where n is the number of words in the corpus. The weakest possible affix removal stemmer would be one that changes no characters in any stemmed word. Such a stemmer would have one word per conflation class. The index compression factor, number of words and stems that differ, mean characters removed, and mean and median modified Hamming distance between word and stem in this case would all be zero.

Stemmer Similarity Metrics

There are several reasons to compare stemmers. For example, information retrieval system designers need to understand the performance of different stemmers in choosing one for index creation. A retrieval system might offer users the choice of stemmers as a way to control recall and precision. A comparison of stemmers might also serve as the basis of developing better stemmers.

We propose a stemmer metric for pairs of stemmers. Currently there is no such metric for stemmer similarity. We suggest a statistic based on the inverse of the modified Hamming distance between stems produced by pairs of stemmers as our stemming similarity metric.

It is easy to confuse stemmer strength and similarity metrics. A stemmer strength metric maps a stemmer to a number that indicates how much a stemmer changes words to produce stems. A stemmer similarity metric maps n-tuples of stemmers (n at least 2), to a number indicating stemmer likeness. Note that a stemmer strength metric is a function of one stemmer, while a stemmer similarity metric is a function of at least 2 stemmers.

The stemmer similarity metric M for a pair of stemming algorithms A1 and A2, given a wordlist W, is the inverse of the mean modified Hamming distance (d) for all words in the wordlist, $M(A1,A2,W) = n/\Sigma\, d(x_i,y_i)$, for i ranging from 1 to n, where n is the size of W and for all words w_i in W, x_i is the result of the application of A1 to w_i and y_i is the is the result of the application of A2 to w_i. The inverse of the mean is used so that more similar algorithms will have higher values of M.

For example, suppose W = {brittle, engineered, fairies} and that stemming algorithm A1 produces the stems {brit, engineer, fairy} from W, and stemming algorithm A2 produces {britt, engineered, fairi} from W. Then M(A1,A2,W) is computed by dividing the wordlist size (3) by the sum of the modified Hamming distances between the stems produced by each stemmer for each word (for example, the modified Hamming distance between brit and britt is 1). The result is 3/(1+2+1) = 0.75 as our measure of the similarity of algorithms A1 and A2.

We use this similarity metric to make pair wise comparisons of the four stemmers. A metric is said to be valid if it measures what it is supposed to measure. The validity of a metric is usually established by showing that the proposed metric is correlated with other known measures of the concept. We establish the validity of M by showing that it agrees with the relative strengths of pairs of stemmers. For example, pairs of stemmers most alike in strength should have the highest similarity measure, and pairs of stemmers least alike in strength should have the lowest similarity measure.

The Wordlist

Grady Ward's Moby project has collected large word lists from a variety of sources (ftp://sable.ox.ac.uk/pub/wordlists/Moby). We took the Moby Common Dictionary wordlist, consisting of 74,550 words appearing in at least two published dictionaries, and combined it with the 20,046 word UNIX spelling dictionary. We then removed all entries not consisting entirely of lower case letters, thus eliminating proper names, abbreviations, hyphenated terms, etc. The resulting wordlist is 49,656 common English words with an average length of 8.07 letters, and a standard deviation of 2.53 letters.

The wordlist used for this study includes all simple, common words that several groups of dictionary writers have found to be in wide use among native American English speakers. This wordlist is a representative sample of words that stemmers will be applied to in practice.

Descriptive Stemmer Data

The four algorithms in this study, Lovins [1968], Paice

	Lovins	Paice	Porter	S-removal
Mean	1.72	1.98	1.16	0.03
Stndrd Dev.	1.64	1.92	1.40	0.19
Minimum	0	0	0	0
25th Percentile	0	0	0	0
Median	1	2	1	0
75th Percentile	3	3	2	0
Maximum	10	13	9	3

Table 1: Modified Hamming Distance Descriptive Statistics

[1990], Porter [1980], and S-removal [Harman 1991], are all longest-match affix removal algorithms. All are rule-driven and run in linear time proportional to the size of their input.

Table 1 shows summary descriptive statistics for the results of running the four stemmers on our wordlist. All data used are modified Hamming distances between words and their stems. The distributions are hyperbolic for all four data sets. The maximum Hamming distance is 13 and the mode is 0 for all stemmers (in other words, most often the stemmers do not change the word).

The statistics in Table 1 show that the stemmers differ in strength, and that some pairs are more similar than others. The analyses that follow are used to establish the relative strengths and similarities of the stemmers.

Stemmer Strength Comparisons

In this section we analyze the strengths of the four stemmers. We first consider the hypothesis that the stemmers are all of equal strength. To test this hypothesis, we conducted a pair wise sign test of modified Hamming distances for the four stemmers. The sign tests showed that there are significant differences between all pairs of stemmers ($p<0.0001$).

We looked at the six measures of stemmer strength described above; the results are displayed in Table 2.

The six metrics are remarkably consistent in ranking the relative strengths of the stemmers. Except for the median modified Hamming distance metric, which shows a tie between the Lovins and Porter stemmers, each metric places the stemmers in the following order, from strongest to weakest: Paice, Lovins, Porter, and S-Removal.

If a single numeric stemmer strength metric is desired, we recommend the mean modified Hamming distance metric for the following reasons:

- The mean modified Hamming distance metric discriminates between stemmers that the median modified Hamming distance metric does not, making it preferable as a tool for comparing stemmer strength.
- The mean modified Hamming distance metric accounts for ending character transformations while the mean characters removed metric does not, making the former a more sensitive measure.
- Both the compression factor and the mean conflation class size are indirect measures of stemmer strength in the sense that they reflect the effects on other entities (the inverted index and the set of conflation classes) rather than the results of the stemmer itself on terms. The mean modified Hamming distance is a direct measure and should be less sensitive to external factors that might affect these other two metrics.
- The count of the number of words in a corpus for which the word and stem differ is a gross measure of stemmer strength because it does not account for how different each word is from its stem, only that it is different. The mean modified Hamming distance metric reflects the effect of a stemming algorithm on each letter of every word.

Based on these measures, we found that Paice is the strongest stemmer, with Lovins somewhat weaker but still quite strong. Porter is considerably weaker than both Paice and Lovins, but in turn is much stronger than S-Removal, which in comparison with the rest is a very weak stemmer. These measures of stemmer

Stemmer	**Mean Modified Hamming Distance**	**Median Modified Hamming Distance**	**Mean Characters Removed**	Compression Factor	Mean Conflation Class Size	Word and Stem Different
Lovins	1.72	1	1.67	0.29	1.42	34437 (69.4%)
Paice	1.98	2	1.94	0.33	1.49	34533 (69.5%)
Porter	1.16	1	1.08	0.17	1.20	27897 (56.2%)
S-Removal	0.03	0	0.03	0.01	1.01	1636 (3.3%)

Table 2: Modified Hamming Distance Descriptive Statistics

strength correspond exactly with the mean modified Hamming distance metric.

Stemmer Similarity

The purpose of stemmer similarity analysis is to show the degree of relatedness of pairs of stemmers. This helps with picking a stemmer, and with the evaluation of new stemmers. Selection of an existing stemmer should be based on its uniqueness and its strength (discussed above). When a new stemmer is proposed, it can be tested against existing ones using a wordlist. If it is nearly identical to an existing stemmer, then arguments for its adoption will need to be made on the grounds of greater efficiency or simplicity.

Table 3 shows the stemmer similarity metric M and the percentage of stems the same. This data suggest that the stemmer similarity pairings from most to least similar are: Paice-Lovins, Porter-Lovins, Porter-Paice, S-removal-Porter, S-removal-Lovins, and S-removal-Paice.

Pairs		Percentage of Stems the Same
Paice-Lovins	1.306	59.8
Porter-Lovins	1.248	57.5
Porter-Paice	1.041	53.7
S-Removal-Porter	0.886	46.0
S-Removal-Lovins	0.592	31.5
S-Removal-Paice	0.515	31.2

Table 3: Similarity Measure M and Percentage of Stems the Same

The validity of M is supported by the fact that M rates stemmer likeness in perfect agreement with the relative strength of pairs of stemmers. Specifically the Paice stemmer is clearly the strongest of the four stemmers, followed closely in strength by the Lovins stemmer. The Porter stemmer is weaker than Lovins and Paice, and the S-removal stemmer is far weaker than the others. The metric M rates the Lovins and Paice stemmers as the most similar, as one would expect given their similar strength. The Porter stemmer is rated as of intermediate similarity to Lovins and Paice. The S-removal stemmer is rated as the least similar to any of the others. The close correspondence between values of M and the likeness of stemmers derivable from their relative strengths validates M as a measure of stemmer similarity.

Outlier Analysis

Boxplots such as those in Figure 1 graphically represent distributions. The top of the box is the 75th percentile, the bottom of the box is the 25th percentile, and the bar in the middle of the box is the median. The lines extending from the box are called whiskers, and the midspread is the span between the 25th and 75th percentiles of the distribution. The ends of the whiskers are 1.5 midspreads from the top and bottom of the box, respectively. Outliers are data points lying beyond 1.5 midspreads from the median of a distribution. These are indicated as circles in the diagram. Extreme outliers, represented by asterisks, lie at least 2 midspreads above the 75th percentile or below the 25th percentile. The distribution of Hamming distances between words and their stems shows both outliers in all the stemmer distributions and extreme outliers in the Paice, Porter, and S removal stemmer distributions. Examination of these outliers leads to some interesting observations.

Outliers are unusual values of one sort or another, and in this case many of the outliers are clearly over-stemmed (too many characters were removed). For example, the Paice stemmer reduces "ultranationalism" to "ultra." Outlier analysis thus leads to data helpful for improving the stemmers.

Over-stemming is correlated with stemmer strength. In particular, over-stemming is very characteristic of the Paice stemmer, frequently occurs with the Lovins stemmer, is relatively infrequent with the Porter stemmer, and did not occur with the S removal stemmer. Such stemming mistakes may have an effect on recall and precision.

Over stemming also seemed to be most prevalent in words where the root morphemes are in the middle of the word. For example, "ultranationalism" has the root morpheme ("nation") between a prefix and two suffixes. The outliers were often over-stemmed all the way back to the prefix, or at least into the root morpheme. This suggests that prefix removal done in conjunction with suffix removal might produce better

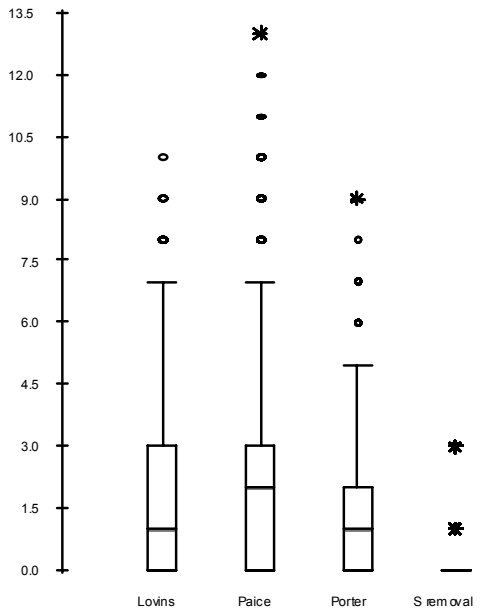

Figure 1: Boxplots of Hamming Distances for four Stemmers

stemming results.

One way to improve stemmer performance (though not explored in the literature) is to keep lists of stemmer exceptions—words for which a stemmer fails. The exception list can be searched before stemming, and if the word is present, its correct stem can be returned. The outliers provide many instances of words that might be included in such a stemmer exception list.

Summary

This study evaluated the strength of, and similarity among, four affix removal stemming algorithms: Lovins, Paice, Porter, and S-Removal. Strength and similarity were evaluated in different ways, including new metrics based on the Hamming distance measure. Data was collected on stemmer outputs for a list of 49,656 English words derived from the UNIX spelling dictionary and the Moby corpus.

Stemmer strength, that is, how much a stemmer changes words that it stems, is important because it is predictive of recall and precision and of index compression. A stronger stemmer will, on average, increase recall, decrease precision, and increase index compression. We tested the hypothesis that the stemmers are all of equal strength using a pair wise sign test of modified Hamming distances for the four stemmers. The sign tests showed that there are significant differences between all pairs of stemmers ($p<0.0001$).

We considered six metrics for evaluating stemmer strength. The metrics we considered are consistent in ranking the relative strengths of the four stemmers examined in this study in the following order, from strongest to weakest: Paice, Lovins, Porter, and S-Removal.

Stemmer similarity analysis is intended to show the degree of relatedness of pairs of stemmers. This can help with selecting among stemmers and with the evaluation of new stemmers. Based on the stemmer similarity metric M (inverse modified Hamming distance) the pairings from most to least similar are: Paice-Lovins, Porter-Lovins, Porter-Paice, S-Removal-Porter, S-Removal-Lovins, and S-Removal-Paice.

We examined outliers in the distribution of modified Hamming distances between words and their stems for each stemmer. Those cases where a stemmer removed an unusually large number of characters from a word were often clear instances of over-stemming (too many characters were removed). Over-stemming is correlated with stemmer strength. In particular, over-stemming is characteristic of the Paice stemmer, frequently occurs with the Lovins stemmer, is relatively infrequent with the Porter stemmer, and did not occur with the S-removal stemmer.

Note

This paper is a summary of a technical report of the same name from the Software Engineering Guild, 1999.

References

Frakes, W. "Stemming Algorithms." In Frakes, W. and R. Baeza-Yates, (ed.) Information Retrieval: Data Structures and Algorithms. Englewood Cliffs, NJ: Prentice-Hall, 1992.

Harman, D. "How Effective is Suffixing." Journal of the American Society for Information Science 42 (1), 1991, 7-15.

Lovins, J. B. "Development of a Stemming Algorithm." Mechanical Translation and Computational Linguistics 11, 1968, 22-31.

Paice, Chris D. "Another Stemmer." SIGIR Forum 24 (3), 1990, 56-61.

Porter, M. F. "An Algorithm for Suffix Stripping." Program 14, 1980, 130-137.