# One Term or Two?

Kenneth Ward Church
AT&T Bell Laboratories
Murray Hill, NJ, USA 07974
kwc@research.att.com

**Abstract**

How effective is stemming? Text normalization? Stemming experiments test two hypotheses: one term (+stemmer) or two (–stemmer). The truth lies somewhere in between. The correlations, ρ, between a word and its variants (e.g., $+s$, $+ly$, +uppercase) tend to be small (refuting the one term hypothesis), but non-negligible (refuting the two term hypothesis). Moreover, ρ varies systematically depending on the words involved; it is relatively large for a good keyword, $\rho(hostage, hostages) \approx 0.5$, and small for pairs with little content, $\rho(anytime, Anytime) \approx 0$, or conflicting content, $\rho(continental, Continental) \approx 0$.

## 1. How effective is suffixing? Text normalization? NLP?

Many systems use a stemmer to map morphological variants, e.g., *hostage* and *hostages*, into a single term. Do stemmers help retrieval performance? Frakes (1992, table 8.1, p. 148) summarizes a number of stemming experiments, many of which failed to find much of a difference in terms of precision and recall (though there have been a few counter-examples such as Krovetz (1993)):

> ‘‘For none of the collections is the improvement of one method over the other really dramatic, so that in practice either procedure might reasonably be used.’’ (Salton and Lesk, 1968, p. 28).

> ‘‘Although individual queries were affected by stemming, the number of queries with improved performance tended to equal the number with poorer performance, thereby resulting in little overall change for the entire test collection.’’ (Harman, 1991, pp. 13-14)

These results are disturbing for those of us working in natural language processing (NLP). If it is hard to show that something as simple as stemming is helpful, how can we possibly justify our interests in more challenging forms of natural language processing such as part of speech tagging, word sense disambiguation, synonymy, phrase identification and parsing?

## 2. One term or two?

Most stemming experiments consider just two conditions:

1. +Stemmer: treat morphological variants as the *same* term, and
2. –Stemmer: treat morphological variants as *different* terms.

Roughly speaking, the two conditions correspond to assuming that the correlations, ρ, among the variant forms are either huge or negligible.

## 3. Estimating Correlations among Variant Forms

ρ is estimated from the four cells of the contingency matrix, *a*, *b*, *c*, *d*, as illustrated in Table 1. The four cells show the number of documents in a corpus of 1988 Associated Press (AP) articles that contain both *hostage* and *hostages* ($a$), the first and not the second ($b$), the second and not the first ($c$), and neither ($d$). The total number of documents in the collection is: $D = a + b + c + d$.

Table 1: A Contingency Table

| | hostages | |
|---|---|---|
| hostage | 619 ($a$) | 479 ($b$) |
| | 648 ($c$) | 78,223 ($d$) |

Let a document be represented as *x*. Each of the elements, $x_i$, is a binary variable indicating the presence or absence of the $i^{th}$ term. In other words, *x* is a ‘‘bag of words’’ with no frequencies. We estimate the joint probability, $Pr(x_i = 1 \& x_j = 1)$, with $a/D$, and the marginal probabilities, $Pr(x_i = 1)$ and $Pr(x_j = 1)$, with $(a+b)/D$ and $(a+c)/D$, respectively. $\sigma_i^2$ and $\sigma_j^2$, the estimates of the variance over documents for the $i^{th}$ and $j^{th}$ terms, are $\frac{a+b}{D} - (\frac{a+b}{D})^2$ and $\frac{a+c}{D} - (\frac{a+c}{D})^2$, respectively. The correlation, ρ, is the difference between the joint probability and chance, normalized appropriately by the variances so that $-1 \le \rho \le 1$.

$$\rho_{i,j} = \frac{Pr(x_i = 1 \& x_j = 1) - Pr(x_i = 1)\,Pr(x_j = 1)}{\sigma_i \sigma_j}$$

When $\rho \approx 1$, the two forms, *i* and *j*, count as a single term; the presence or absence of one in a document gives us no additional information over what we know from looking at the other. Conversely, when $\rho \approx 0$, the two forms count as two terms; the presence or absence of one form tells us little or nothing about the presence or absence of the other. We rarely find negative correlations, but if we did, the presence of one form would predict the absence of the other, and vice versa.

### 4. The Bahadur and Lazarsfeld (BL) Expansion

We suggested above that stemming is similar to assuming $\rho \approx 1$, and that the alternative is similar to assuming $\rho \approx 0$. We can make this statement precise in terms of the Bahadur and Lazarsfeld (BL) expansion (Duda and Hart, 1973, pp. 111-113), (Salton, 1989, pp. 345-349). Let the probability of a document from the relevant set be:

$$Pr(x|rel) = \prod_{k=1}^{t} p_k^{x_k} (1 - p_k)^{1 - x_k} [1 + A]$$

where $p_i$ is $Pr(x_i = 1 | rel)$, and $A$ is a correction factor that accounts for the correlations among terms, $\rho$.

$$A = \sum_{i<j} \rho_{i,j} \delta_i \delta_j + \sum_{i<j<k} \rho_{i,j,k} \delta_i \delta_j \delta_k + \cdots$$

$\delta_i$ is a normalization of $x_i$ (that subtracts the mean and divides by the standard deviation).

$$\delta_i = \frac{x_i - p_i}{\sqrt{p_i (1 - p_i)}}$$

In practice, it is necessary to introduce various approximations, e.g., setting many of the correlations to 0, stemming, text normalization, etc. There are too many correlations in the BL expansion, many of which are hard to estimate because of sparse data, and probably make little difference in terms of precision/recall. Even though we might not be able to afford to work with the BL expansion directly in a practical setting, it is useful, nevertheless, as a theoretical device to help us better understand the consequences of certain popular approximations.

Consider the following simple example of the stemming approximation. Suppose that the vocabulary consists of just two words, *hostage* and *hostages*, and that both words have the same probability, $p_{rel}$, of being found in a relevant document. Imagine a document that contains both *hostage* and *hostages*. This document should have a probability of $p_{rel}$ or $p_{rel}^2$, depending on whether we treat *hostage* and *hostages* as one term or two.

As least in this simple example, we can show that the two conditions correspond to assuming that $\rho$ is either 0 or 1. That is, $Pr(x|rel) = p_{rel}^2 [1 + A]$ is either $p_{rel}^2$ or $p_{rel}$, depending on whether $\rho = 0$ or $\rho = 1$.[1] In general, of course, the relationship between $\rho$ and stemming can be considerably more complicated.

### 5. A term and a half?

Empirically, we find that *hostage* and *hostages* should count as more than one term, but less than two. Using the numbers in Table 1, we find that $\rho \approx 0.52$, roughly a term and a half. Note that if $\rho = 0.5$ in the example above, then $Pr(x|rel) = 0.5(p_{rel}^2 + p_{rel})$, halfway between what it was for $\rho = 0$ ($Pr(x|rel) = p_{rel}^2$) and what it was for $\rho = 1$ ($Pr(x|rel) = p_{rel}$).

To get a better sense of the distribution of correlations, we collected a set of 999 morphologically related pairs such as *hostage* and *hostages*.[2] A histogram of the correlations is shown in Figure 1. Most are greater than 0, but far from 1. The correlations are small but non-negligible. It would be a mistake to ignore these correlations (–stemmer), but it would be even more of a mistake to assume that they are nearly perfect (+stemmer). It is not surprising that stemming helps little (if any). In fact, we might even expect stemming to degrade performance (though one could imagine all sorts of reasons why experiments have failed to find much of a difference either way).

### 6. Some Pairs Count More Than Others

As we have seen, the correlations are more than 0 (–stemmer) but far from 1 (+stemmer). One could imagine various compromises such as $\rho \approx 0.5$ (a term and a half) or

---

1. Since $x_i = x_j = 1$ and $p_i = p_j = p_{rel}$, $\delta_i = \delta_j = \frac{1 - p_{rel}}{\sqrt{p_{rel}(1 - p_{rel})}}$. Because there are just two words in the vocabulary, $A = \rho \delta_i \delta_j = \rho (\frac{1 - p_{rel}}{\sqrt{p_{rel}(1 - p_{rel})}})^2 = \rho (\frac{1}{p_{rel}} - 1)$. If $\rho = 0$, then $1 + A = 1$ and consequently, $Pr(x|rel) = p_{rel}^2 [1 + A] = p_{rel}^2$. If $\rho = 1$, then $1 + A = p_{rel}^{-1}$ and therefore, $Pr(x|rel) = p_{rel}^2 [1 + A] = p_{rel}$.

2. Hopefully the results of the experiment should not depend too much on the details of the sampling procedure. We collected the set of 999 morphologically related pairs by extracting the first 5000 distinct words from a collection of Associated Press (AP) newswire (written during 1988), and then arbitrarily appending an *s* to the end of each of these 5000 words. This process generated large numbers of spurious pairs such as (*the*, *the* + *s*). In addition, we were concerned that the estimate of $\rho$ could be unstable if either member of the pair had a small document frequency. To exclude spurious and/or troublesome pairs from the sample, we sorted the 5000 pairs by the document frequency of the $+s$ form, and selected the top 1000. We then checked the pairs by hand to verify that they seemed reasonable. One pair, (*countrie, countries*), was thrown out because the baseform appeared to be a typographical error.

ρ≈0.1, but Table 2 (above) and Figure 2 (below) suggest that ρ should not be modeled as a constant because it depends on the words involved.

The estimates of the correlations in Figure 1 were repeated over a second dataset of similar material, a corpus of 1989 AP articles. The ten pairs with the largest correlations in Figure 1 and the ten pairs with the smallest correlations in Figure 1 are shown in Table 2. The size of the correlation is remarkably consistent across datasets. All 999 pairs are shown in Figure 2. The large correlation of correlations (0.94) indicates that ρ varies systematically by word.

Intuitively, pairs with larger correlations in Table 2 appear to be more useful for retrieval purposes than pairs with smaller correlations. If an AP story mentions the word *hostage*, for example, then it is probably about hostages. In contrast, if an AP story mentions the word *await*, it could be about practically anything. We will return to this conjecture in Section 10.

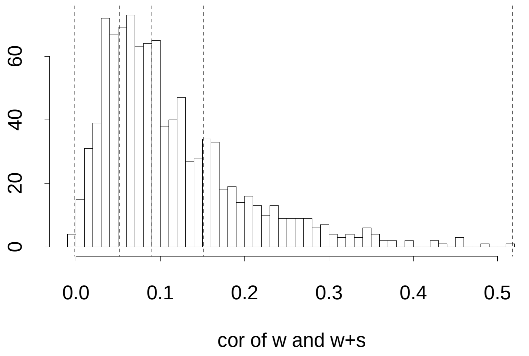


Figure 1: The correlations of 999 morphologically related forms (w and w+s) are systematically positive, but far from 1. Only 4 of the 999 correlations were negative. The quantiles (min, 25%, median, 75%, max) are indicated by dashed lines at: –0.002, 0.052, 0.090, 0.151, 0.518. The ten largest and the ten smallest correlations are shown in Table 2.

Table 2: Some words have large correlations, and some don't

| Large Correlations | | | Small Correlations | | |
|---|---|---|---|---|---|
| 1988 | 1989 | | 1988 | 1989 | |
| 0.52 | 0.50 | hostage(s) | 0.00 | 0.01 | await(s) |
| 0.49 | 0.44 | reactor(s) | 0.00 | 0.01 | ground(s) |
| 0.46 | 0.52 | rebel(s) | 0.00 | 0.01 | possession(s) |
| 0.46 | 0.50 | guerrilla(s) | 0.00 | 0.04 | boast(s) |
| 0.46 | 0.65 | abortion(s) | 0.00 | 0.02 | belonging(s) |
| 0.44 | 0.18 | delegate(s) | 0.00 | 0.01 | compare(s) |
| 0.43 | 0.44 | drug(s) | 0.00 | 0.01 | direct(s) |
| 0.42 | 0.44 | stock(s) | 0.00 | 0.04 | shield(s) |
| 0.40 | 0.34 | pesticide(s) | 0.00 | 0.01 | last(s) |
| 0.39 | 0.41 | airline(s) | 0.01 | 0.02 | urge(s) |

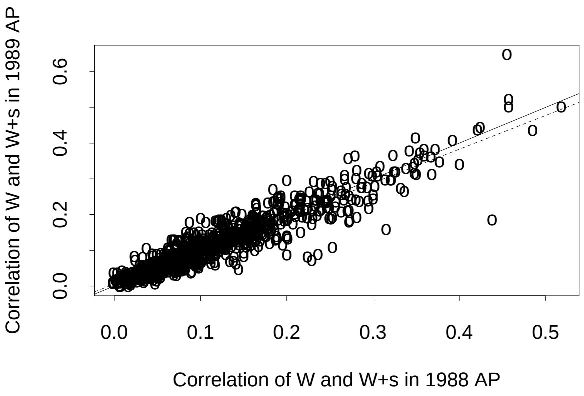


Figure 2: Some terms count more than others. The strength of the correlation between a word and its $+s$ form varies systematically from one pair to the next. The correlations for the 999 pairs in Figure 1 are estimated twice, once from the 1988 AP data and once from the 1989 AP data. The two sets of estimates are highly correlated (0.94). The regression line (dotted line) has a slope of 0.94 and an intercept of 0.0056, close to the main diagonal (solid line).

Table 3: Estimates are Robust

| | 1988 | 1989 | 1990 | 1991 | 1992 | |
|---|---|---|---|---|---|---|
| Better | 0.518 | 0.500 | 0.488 | 0.597 | 0.347 | hostage(s) |
| Key- | 0.485 | 0.435 | 0.337 | 0.407 | 0.478 | reactor(s) |
| words | 0.457 | 0.522 | 0.534 | 0.511 | 0.483 | rebel(s) |
| | 0.457 | 0.501 | 0.434 | 0.436 | 0.426 | guerrilla(s) |
| | 0.455 | 0.648 | 0.512 | 0.546 | 0.513 | abortion(s) |
| | 0.438 | 0.184 | 0.166 | 0.218 | 0.375 | delegate(s) |
| | 0.424 | 0.443 | 0.418 | 0.371 | 0.336 | drug(s) |
| | 0.421 | 0.436 | 0.464 | 0.440 | 0.266 | stock(s) |
| | 0.400 | 0.340 | 0.302 | 0.300 | 0.243 | pesticide(s) |
| | 0.392 | 0.407 | 0.356 | 0.438 | 0.403 | airline(s) |
| Worse | –0.002 | 0.007 | 0.022 | 0.030 | 0.010 | await(s) |
| Key- | –0.002 | 0.009 | 0.000 | 0.003 | 0.004 | ground(s) |
| words | –0.001 | 0.008 | 0.018 | 0.010 | –0.004 | possession(s) |
| | –0.001 | 0.036 | 0.009 | 0.037 | 0.020 | boast(s) |
| | 0.001 | 0.016 | 0.003 | 0.020 | 0.027 | belonging(s) |
| | 0.002 | 0.014 | 0.014 | 0.019 | 0.023 | compare(s) |
| | 0.003 | 0.007 | 0.011 | 0.012 | 0.001 | direct(s) |
| | 0.003 | 0.036 | 0.068 | 0.092 | 0.052 | shield(s) |
| | 0.004 | 0.009 | 0.009 | 0.008 | 0.006 | last(s) |
| | 0.005 | 0.021 | 0.017 | 0.029 | 0.022 | urge(s) |

**7. How reliable are the estimated correlations?**

One way to address this question would be to perform a set of blackbox retrieval experiments and see whether the correlations could be used in the BL expansion (or some other way) to improve precision/recall. But given the history of failure-to-find stemming experiments, we feared that a straightforward experimental design such as this was unlikely to succeed. There are so many factors at work that the morphological effects of interest would probably be swamped out by some other factor that we should have control for, but didn't even know about.

We decided instead to adopt a glassbox design which offers more sensitivity to the factors of interest (but less insight as to how the factors could be exploited in an operational system). The AP news was divided into five datasets, one for each year between 1988 and 1992. The 999 correlations were estimated five times, one for each of the five datasets. If the estimates are reliable, the estimate based on one dataset should be a good predictor of the estimate based on another dataset.

Table 2 and Figure 2 already found a strong agreement between the estimates in 1988 and 1989. The same arguments are used to show agreement among the estimates over all five datasets in Tables 3-4 and Figure 3. Table 3 is analogous to Table 2. The first ten pairs (labeled ''better keywords'') have consistently larger correlations in all five datasets than the second ten pairs (labeled ''worse keywords''). Figure 3 shows 20 scatter plots like the one in Figure 2. Each scatter plot compares the estimates for all 999 pairs in two different datasets. The scatter plot in the upper-right corner, for example, compares estimates based on 1988 AP articles with the same estimates based on 1992 AP articles. Most of the points in the scatter plots fall near the main diagonal. This observation is made more precise in Table 4, which summarizes each scatter plot with a single number, a correlation indicating how well an estimate based on one dataset can be used to predict an estimate based on another dataset. The large correlations in Table 4 indicate that the estimates are highly reliable.

As a side note, we observe that the correlations in Table 4 are slightly larger near the main diagonal, indicating that there is a tendency for the estimates to degrade over time. If you want to predict the correlation between *hostage* and *hostages* in this year's AP, it is better to use last year's estimate than an estimate from ten years ago.

**8. +ly Morphology**

It has been previously suggested that $+s$ affixes are somehow special. Salton and Lesk (1968, p. 28), for example, distinguished pairs like *apple* and *apples* from *analyzer* and *analyzing*. It is not clear why $+s$ morphology should be special, but one possibility is that $+s$ morphology is often associated with nouns, and that nouns might be relatively good keywords on average compared with other parts of speech.

We decided to compare $+s$ morphology with $+ly$ morphology: *swift* $\rightarrow$ *swiftly*. $+ly$ turns adjectives into adverbs, and generally does not apply to nouns (though some stemmers might analyze *godly* and a few others as noun+ly). We expected that $+ly$ pairs would have relatively small correlations compared to $+s$ pairs. This prediction is confirmed in Figure 4, though we should be careful about comparisons across different sets of words. It would be interesting to look at a few more morphological processes and see if it is true that processes that operate on nouns display larger correlations than those that operate on other parts of speech.

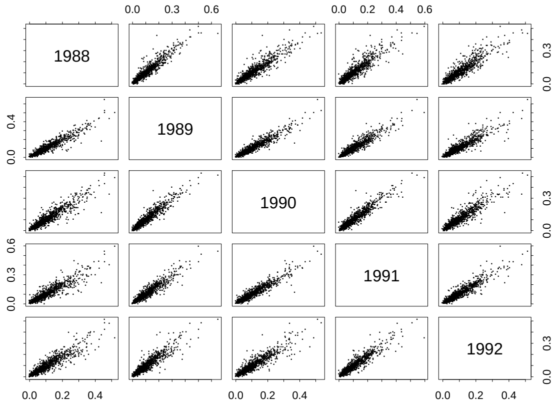


Figure 3: The estimates are highly reliable across datasets. Each point shows the estimated correlation between a word and its +s variant in two different years of AP articles. Most points are near the main diagonal, indicating that estimates based on one dataset are quite similar to estimates based on another.

Table 4: Correlations of correlations in Figure 3

| | 1988 | 1989 | 1990 | 1991 | 1992 |
|---|---|---|---|---|---|
| 1988 | | 0.94 | 0.92 | 0.91 | 0.91 |
| 1989 | 0.94 | | 0.95 | 0.94 | 0.91 |
| 1990 | 0.92 | 0.95 | | 0.94 | 0.91 |
| 1991 | 0.91 | 0.94 | 0.94 | | 0.93 |
| 1992 | 0.91 | 0.91 | 0.91 | 0.93 | |

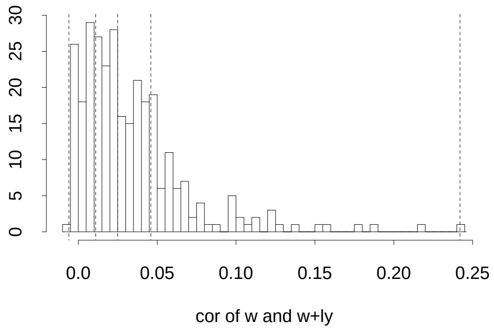


Figure 4: The correlation of a word and its $+ly$ form is also systematically positive, but smaller than $+s$. 45 of the 300 pairs had negative correlations. The quantiles are indicated by dashed lines at: –0.006, 0.011, 0.025, 0.046, 0.242. The ten largest and the ten smallest correlations are shown in Table 5.

Table 5: +ly correlations

| Better Keywords | | Worse Keywords | |
|---|---|---|---|
| 0.242 | sexual(ly) | –0.006 | New(ly) |
| 0.220 | racial(ly) | –0.005 | hot(ly) |
| 0.189 | quarter(ly) | –0.004 | sole(ly) |
| 0.176 | mental(ly) | –0.003 | Week(ly) |
| 0.159 | alleged(ly) | –0.003 | disorder(ly) |
| 0.153 | illegal(ly) | –0.003 | sound(ly) |
| 0.136 | Night(ly) | –0.003 | square(ly) |
| 0.126 | environmental(ly) | –0.002 | objective(ly) |
| 0.122 | sharp(ly) | –0.002 | practical(ly) |
| 0.121 | political(ly) | –0.002 | swift(ly) |

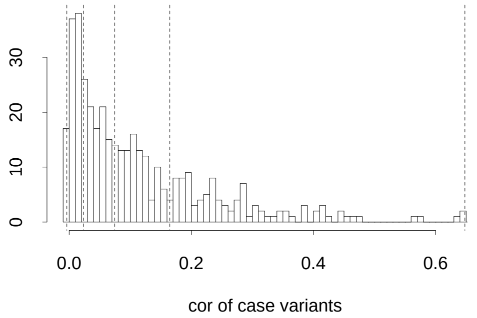


Figure 5: The correlations of case variants are also systematically positive, but far from 1. Only 15 of the 394 correlations were negative. The quantiles (min, 25%, median, 75%, max) are indicated by dashed lines at: –0.004, 0.023, 0.075, 0.165, 0.648. The ten largest and the ten smallest correlations are shown in Table 6.

## 9. Upper and Lower Case

Case variants such as *hurricane* and *Hurricane* are similar to morphological variants. Retrieval systems are even more likely to ‘‘normalize’’ case variants, and yet, Figure 5 shows that their correlations are no larger than the correlations of the $+s$ pairs shown in Figure 1. If the stemming experiments referred to in Section 1 had investigated case variants, they probably would have found that the standard practice of normalizing case is no better (and possibly worse) than stemming.

Table 6 is analogous to Table 3. Table 6 shows the ten best pairs and the ten worst pairs in Figure 5. Sometimes case variants refer to the same thing (e.g., *hurricane* and *Hurricane*), sometimes they refer to different things (e.g., *continental* and *Continental*), and sometimes they don’t refer to much of anything (e.g., *anytime* and *Anytime*). The correlations are relatively large in the first case, and small in the second and third.

Figure 6 and Table 7 are analogous to Figure 3 and Table 4. Figure 6 and Table 7 show that the correlations of case variants are remarkably stable over time, just as we saw for the correlations of case variants in Figure 3 and Table 4.

As before, the correlations in Table 7 are slightly larger near the main diagonal, indicating a tendency for the correlations to degrade over time.

Table 6: Case Variants

| | 1988 | 1989 | 1990 | 1991 | 1992 | |
|---|---|---|---|---|---|---|
| Good | 0.65 | 0.60 | 0.58 | 0.57 | 0.64 | [Hh]urricane |
| key- | 0.64 | 0.62 | 0.59 | 0.57 | 0.57 | [Pp]ope |
| words | 0.64 | 0.47 | 0.48 | 0.29 | 0.39 | [Ee]mperor |
| with | 0.58 | 0.54 | 0.52 | 0.59 | 0.51 | [Ll]ottery |
| com- | 0.57 | 0.49 | 0.53 | 0.60 | 0.48 | [Zz]oo |
| mon | 0.47 | 0.49 | 0.47 | 0.43 | 0.54 | [Bb]allet |
| mean- | 0.46 | 0.24 | 0.44 | 0.37 | 0.19 | [Gg]ulf |
| ing | 0.45 | 0.48 | 0.35 | 0.15 | 0.21 | [Cc]anal |
| | 0.44 | 0.46 | 0.35 | 0.35 | 0.37 | [Ii]mmigration |
| | 0.44 | 0.45 | 0.48 | 0.46 | 0.42 | [Mm]useum |
| Little | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | [Tt]roy |
| con- | 0.00 | –0.01 | 0.01 | 0.00 | 0.00 | [Pp]ath |
| tent | 0.00 | –0.01 | 0.00 | 0.00 | 0.01 | [Ee]ditions |
| or | 0.00 | 0.01 | 0.01 | 0.02 | 0.00 | [Cc]ontinental |
| con- | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 | [Bb]urns |
| flict- | 0.00 | –0.00 | 0.00 | 0.00 | 0.00 | [Ll]evy |
| ing | 0.00 | –0.01 | –0.01 | –0.01 | –0.02 | [Aa]dv |
| con- | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | [Hh]aven |
| tent | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | [Rr]ush |
| | 0.00 | 0.02 | 0.00 | 0.00 | 0.01 | [Aa]nytime |

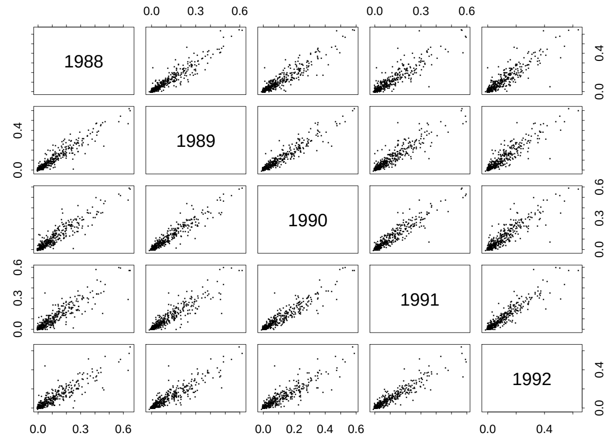


Figure 6: The correlations of case variants are also remarkably stable over time. Each point shows the correlation between an upper case word and its lower case variant in two different years of AP articles.

Table 7: Correlations of correlations in Figure 6

| | 1988 | 1989 | 1990 | 1991 | 1992 |
|---|---|---|---|---|---|
| 1988 | | 0.94 | 0.93 | 0.90 | 0.90 |
| 1989 | 0.94 | | 0.95 | 0.91 | 0.91 |
| 1990 | 0.93 | 0.95 | | 0.94 | 0.90 |
| 1991 | 0.90 | 0.91 | 0.94 | | 0.93 |
| 1992 | 0.90 | 0.91 | 0.90 | 0.93 | |

## 10. Repeated Concepts

Thus far, we have been concerned with estimating the strength of the dependency between two variants of the same thing, either morphological variants or case variants. In some sense, these are all special cases of a general problem of modeling repetition. Standard independence assumptions, such as setting a correlation to zero or imposing a Poisson assumption, fail to account for the fact that whatever the document is about is likely to be repeated more than would be expected by chance.

Consider the probability that a word such as *hostages* will be repeated verbatim in a single document. Under standard independence assumptions, it is extremely unlikely that lightning would strike twice (or half a dozen times) in the same document. But text is more like a contagious disease than lightning. If we see one instance of a contagious disease such as tuberculosis in a city, then we shouldn't be surprised to find quite of few more. Similarly, if we see one instance of *hostages* in a document, we shouldn't be surprised to find more instances of the same word, as well as its variant forms.

Figure 7 shows that the distribution of *hostages* is far from Poisson. It would be extremely surprising to find half a dozen or more instances of *hostages* in a single document under a Poisson model. If we looked at a million years of AP news, we should expect to find just one such document ($\sum_{k \geq 6} D\pi(\theta,k) \approx 10^{-6}$, with $\theta \approx 0.046$ and $D \approx 80,000$), and yet we found dozens in each of the five years under investigation. It has been our experience that the deviations from Poisson tend to be particularly noticeable for good keywords like *hostages*, and less noticeable for crummy keywords like *awaits*.

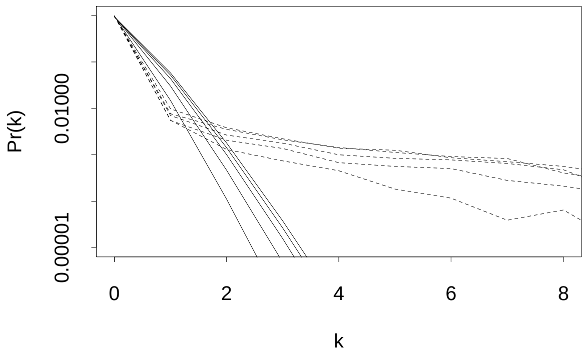


Figure 7: The distribution of *hostages* in the AP is far from Poisson. The dotted lines show the fraction of documents containing $k$ instances of *hostages* in each of the five years between 1988 and 1992. The solid lines show five Poissons, $Pr_P(k) = \pi(\theta,k) = \frac{e^{-\theta}\,\theta^k}{k!}$, where $\theta$ is estimated by the average number of *hostages* per document in each of the five years.

The Two-Poisson Model (Bookstein and Swanson, 1974; Harter, 1975) has been used in IR to model deviations from Poisson. Unfortunately, two Poissons are probably not enough, as illustrated by the dip between the two Poissons in Figure 8. Bookstein and Swanson (1974, p. 317) came to the same conclusion, and suggested a Three Poisson model, though they noted that it would require even more parameters than the Two Poisson model. They then suggested the Negative Binomial following Mosteller and Wallace (1964), which can be viewed as a mixture of infinitely many Poissons. Figure 9 shows that the Negative Binomial fits the data better than the Two Poisson. Katz' K-mixture (Figure 10) fits about as well as the Negative Binomial, but is easier to work with.

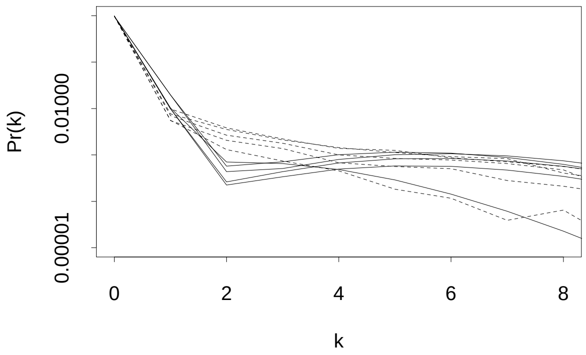


Figure 8: The Two-Poisson Model (solid lines), $Pr_{2P}(k) = \alpha\ \pi(\theta_1,k) + (1-\alpha)\ \pi(\theta_2,k)$, fits the *hostages* data in Figure 7 (dotted lines) better than the Poisson. The parameters were fit using the method of moments (Harter, 1975a, p. 202), which equates the first three observed moments (mean, variance and third moment) with their theoretical values, and solves the system of three equations for $\alpha$, $\theta_1$ and $\theta_2$. Unfortunately, the Two-Poisson Model is often bimodal with an embarrassing dip between the two peaks.

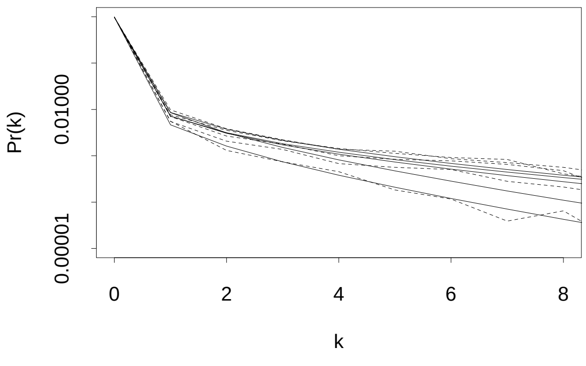


Figure 9: The Negative Binomial (solid lines), $Pr_{NB}(k) = \begin{bmatrix} N+k-1 \\ k \end{bmatrix} P^k\ Q^{-N-k}$, fits the *hostages* data (dotted lines) better than the Poisson and Two Poisson. The two parameters, $N$ and $P$, were fit by method 2 in Johnson and Kotz (1969), which equates the observed mean and IDF with their theoretical values, and solves the system of two equations for $N$ and $P$.

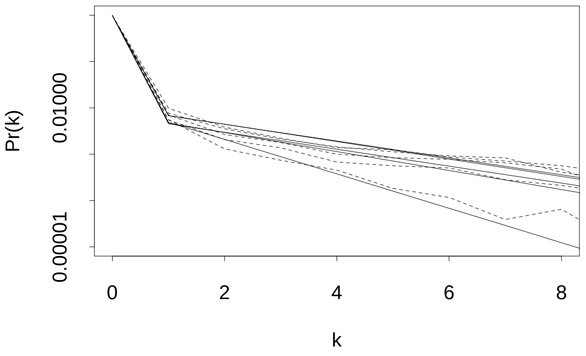


Figure 10: Katz' K-mixture (solid lines), $Pr_K(k) = (1-\alpha)\ \delta_{k,0} + \frac{\alpha}{\beta+1}\ (\frac{\beta}{\beta+1})^k$, fits the *hostages* data (dotted lines) about as well as the Negative Binomial. The two parameters, $\alpha$ and $\beta$, were fit by equating the observed mean ($\hat{\theta}$) and IDF ($\widehat{IDF}$) with their theoretical values, and solving for $\alpha$ and $\beta$: $\beta = \hat{\theta} 2^{IDF} - 1$ and $\alpha = \hat{\theta}/\beta$.

The same words that show a large deviation from Poisson often also exhibit a large correlation with their variant forms. This pattern is illustrated in Tables 8-9 and Figures 11-12, where deviation from Poisson is measured in terms of Residual $IDF = IDF - \widehat{IDF}$. $\widehat{IDF}$ is an estimate of $IDF$ under a Poisson Model: $\widehat{IDF} = -log_2(1 - \hat{\pi}(\theta, 0)) = -log_2(1 - e^{-\theta})$, where $\theta$ is approximated by $\hat{\theta}$, the mean frequency, averaged over documents. The Residual IDFs are shown in Table 8 for the $+s$ form of the twenty pairs in Table 2. The Residual IDFs are large for pairs that had large correlations in Table 2, and near zero for pairs that had near zero correlations in Table 2. Figure 11 shows that the Residual IDF of the $+s$ form can be used to predict the correlation of the morphological pair.

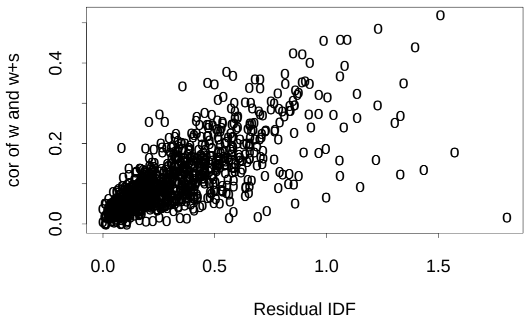


Figure 11: Residual IDF, a measure of deviation from Poisson, can be used to predict the correlation between a word and its morphological variant. Each circle corresponds to one of the 999 pairs in Figure 1. The Residual IDFs were computed over the +s form.

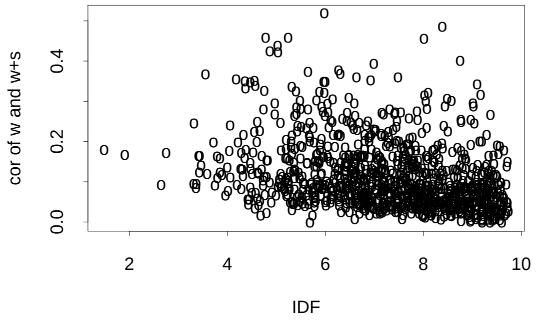


Figure 12: IDF is not as good as Residual IDF for predicting the correlation between a word and its morphological variant.

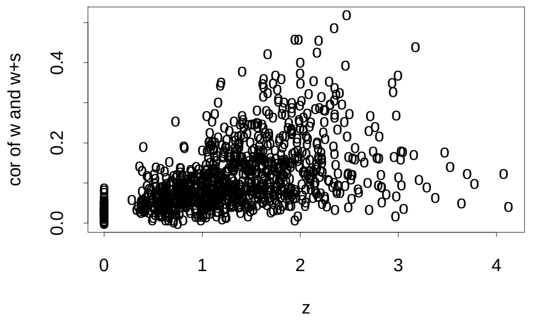


Figure 13: $z = \frac{\theta_1 - \theta_2}{\sqrt{\theta_1 + \theta_2}}$ (Harter, 1975a, p. 204) is better than IDF but not as good as Residual IDF for predicting the correlation between a word and its morphological variant.

Table 8: Residual IDF, IDF and Morphology

| Large Correlations | | | Small Correlations | | |
|---|---|---|---|---|---|
| Resid IDF | IDF | | Resid IDF | IDF | |
| 1.51 | 5.98 | hostages | 0.01 | 9.60 | awaits |
| 1.23 | 8.39 | reactors | 0.11 | 5.68 | grounds |
| 1.09 | 4.78 | rebels | 0.08 | 9.36 | possessions |
| 1.06 | 5.24 | guerrillas | 0.01 | 9.20 | boasts |
| 0.99 | 8.01 | abortions | 0.09 | 8.96 | belongings |
| 1.40 | 5.03 | delegates | 0.01 | 8.77 | compares |
| 0.85 | 4.87 | drugs | 0.05 | 9.19 | directs |
| 0.89 | 5.03 | stocks | 0.10 | 9.40 | shields |
| 0.93 | 8.75 | pesticides | –0.00 | 9.56 | lasts |
| 1.08 | 6.99 | airlines | 0.06 | 8.97 | urges |
| 0 | | *Poisson* | 0 | | *Poisson* |

IDF is not as good as Residual IDF for predicting the correlation of the morphological pair. The correlation in Figure 12 is –0.32, much smaller than the correlation in Figure 11 of 0.73. To the extent that there is a relationship between IDF and the morphological correlation, IDF is *inversely* related to the morphological correlation.

Table 9: Residual IDF is Better than IDF for Predicting the Correlation of a Word and its +s Form

| Dataset | Residual IDF | IDF |
|---|---|---|
| 1988 | 0.73 | –0.32 |
| 1989 | 0.70 | –0.33 |
| 1990 | 0.68 | –0.33 |
| 1991 | 0.70 | –0.34 |
| 1992 | 0.69 | –0.34 |

Many measures of deviation from Poisson have been suggested in the past. Figure 13 shows that $z = \frac{\theta_1 - \theta_2}{\sqrt{\theta_1 + \theta_2}}$ (Harter, 1975a, p. 204), a measure of the separation of the two θs in the Two-Poission, is better than IDF, but not as good as residual IDF. The correlation in Figure 13 is 0.50, more than –0.32 in Figure 12 but less than 0.73 in Figure 11. A large $z$ is often an indicator of a large deviation from Poisson, but not when $\alpha$ is near 0 or near 1.

Why is the deviation from Poisson useful for predicting the correlation of a morphologically related pair? We might speculate that some words have a distribution that is almost like chance, and some words have a distribution that can only be accounted for by positing the existence of certain hidden variables such as relevance, author, genre, etc. The latter set of words are more useful for retrieval purposes because their distribution sheds light on hidden variables of interest.

Standard IDF, on the other hand, is less well suited for identifying hidden variables. If there is a major hostage-taking episode in the news, for example, there are likely to be a relatively large number of articles about *hostages*, and as a result, the IDF for *hostages* might actually go down as the importance of the event goes up, resulting in a negative correlation between IDF and the importance of the keyword.

On this account, a good keyword is one that behaves very differently from chance. This hypothesis runs counter to the standard practice in Information Retrieval of weighting words by IDF, favoring extremely rare words, no matter how they are distributed.

**11. Recommendations**

How can we use these observations in a practical system? In principle, one could estimate the correlations of a large number of morphologically related pairs, and use the estimates in the BL expansion to improve the estimates of $Pr(x | rel)$. It is likely, though, that this approach would fail to produce a significant improvement in precision/recall given the history of negative results.

In the short term, the most productive use of these correlations may be to develop and test new retrieval strategies. We cannot afford to carry out as many full-blown retrieval experiments as we might like to. Retrieval experiments take time; test collections are expensive. And all too often, the experiments fail to find a significant result. We need a cheaper way to test the feasibility of a radical new alternative, and only if it shows promise would we consider a more thorough (and costly) set of retrieval experiments.

The discussion of Residual IDF used the morphological correlations in this way to develop and test a radical alternative to IDF weighting. Rather than weighting words by IDF, favoring extremely rare words no matter how they are distributed, we suggested that words should be weighted by residual IDF, a measure of deviation from Poisson. We developed and tested this hypothesis by using the morphological correlations as a gold standard, and evaluated three weighting schemes, residual IDF, standard IDF and z, on the basis of how well they predicted the standard. This evaluation framework has the advantage that it is relatively easy to collect large quantities of testing and training materials. On the other hand, since there is no guarantee that the morphological constraints reflect the real issues in the retrieval task, we need to check our findings with a set of standard retrieval experiments.

As a side note, we have also used residual IDF to construct a large ‘‘thesaurus’’ of highly correlated pairs such as: *Kurds/Kurdish*, *Estonian/Latvia*, *Boesky/speculator* and *shuttle/Challenger*. It would be prohibitively expensive to compute the correlations for all pairs of words in the AP news, and not very productive because most of the correlations are negligible. We decided to narrow the search to a short list of relatively high grade ore, by requiring that both words have large residual IDFs and that they both appear at least twice in the same document. These steps made it feasible to compute the correlations over the short list, and worth doing so, since most of the correlations that were computed turned out to be significant.

**12. Conclusions**

Stemming and other text normalization procedures are commonly found in many retrieval systems, even though Salton and Lesk (1968), Harman (1991) and others have reported that stemming produces little if any improvement in precision/recall. These experiments are disturbing for those us interested in natural language processing (NLP) techniques such as part of speech, phrase identification, thesauruses and parsing. If something as simple as stemming doesn't help, then how can we justify our interests in more elaborate NLP techniques?

Stemming experiments have tended to consider just two conditions: *hostage* and *hostages* are either one term or two. In fact, the truth lies somewhere in the middle. We estimated the correlations for 999 pairs in five years of Associated Press articles and found that the correlations were almost always larger than 0, but far from 1.

We then considered case variants such as *hurricane* and *Hurricane*, and *continental* and *Continental*. The practice of treating case variants as a single term is just as suspect, if not more so, than the practice of treating morphological variants as a single term. This observation draws into question a whole range of text normalization techniques which have become standard practice. Clearly, we need to be careful about abusing a thesaurus. We need to check the correlations before we blindly replace one synonym with another. But what about less drastic text normalization techniques such as canonicalizing spelling variants such as *IBM* and *I.B.M.* Is this safe? If it isn't, then what is?

The size of the correlation depends on the particular words involved. The correlations are large when both members are good keywords and they both refer to the same thing, e.g., *hostage* and *hostages*. Conversely, they tend to be small when they refer to different things, e.g., *continental* and *Continental*, or when they don't refer to much of anything, e.g., *anytime* and *Anytime*. The correlations tend to be larger for more ''noun-like'' $+s$ pairs, and smaller for less ''noun-like'' $+ly$ pairs. The fact that the correlations vary in systematic ways argues against a simple uniform treatment of the correlations such as either +stemmer (one term) or –stemmer (two terms) or various compromises such as $\rho = 0.5$ (a term and a half) or $\rho = 0.1$.

In some sense, all of these correlations can be viewed as yet another form of repetition. Whatever the document is about is likely to be repeated. As a result, good keywords like *hostage* and its variants are likely to be repeated more than chance, producing large correlations among the variant forms, and large deviations in IDF from what would be expected under a Poisson model of a random keyword with the same word frequency. In contrast, less good keywords like *awaits* are closer to chance in both respects. The distribution of *awaits* is nearly Poisson, and there are almost no correlations between *awaits* and its variants. Deviations from chance are exhibited by large deviations from Poisson and large correlations with related words.

On this account, a good keyword is one that behaves very differently from chance. This hypothesis runs counter to the standard practice in Information Retrieval of weighting words by IDF, favoring extremely rare words, no matter how they are distributed. We suggested that it might be even better to weight words by residual IDF (the difference between the observed IDF and what would be expected under a Poisson model for a random word with comparable frequency), and showed that residual IDF was superior to IDF for predicting morphological correlations.

**Acknowledgments**

This work benefited considerably from extensive discussions with Bill Gale, Slava Katz and David Lewis.

**References**

Bookstein, A., and Swanson, D. (1974) ''Probabilistic Models for Automatic Indexing, *Journal of the American Society for Information Science*, 25:5, pp. 312-318.

Church, K., and Gale, W. (in press) ''Poisson Mixtures,'' *Journal of Natural Language Engineering*.

Duda, R., and Hart, P. (1973) *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York.

Frakes, W. (1992), ''Stemming Algorithms,'' in Frakes, W. and Baeza-Yates, R. (eds.) *Information Retrieval: Data Structures & Algorithms*, Prentice Hall, Englewood Cliffs, N.J.

Harman, D. (1991) ''How Effective Is Suffixing?'' *Journal of the American Society for Information Science*, 42(1):7-15.

Harter, S. (1975a), ''A Probabilistic Approach to Automatic Keyword Indexing: Part I. On the Distribution of Specialty Words in a Technical Literature,'' *Journal of the American Society for Information Science*, 26(4), 197-206.

Harter, S. (1975b), ''A Probabilistic Approach to Automatic Keyword Indexing: Part II. An Algorithm for Probabilistic Indexing,'' *Journal of the American Society for Information Science*, pp. 280-289.

Katz, S. (personal communication).

Krovetz, R. (1993) ''Viewing Morphology as an Inference Process,'' *SIGIR*, pp. 191-202.

Johnson, N., and Kotz, S. (1969) *Discrete Distributions*, Houghton Mifflin, Boston.

Mosteller, F., and Wallace, D. (1964) *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, Massachusetts.

Salton, G., and Lesk, M. (1968) ''Computer Evaluation of Indexing and Text Processing,'' *Journal of the Association for Computing Machinery*, 15:1, pp. 8-36.