On Log-Likelihood-Ratios and the Significance of Rare Events*

Robert C. MOORE
Microsoft Research
One Microsoft Way
Redmond, WA 90052
USA
bobmoore@microsoft.com

Abstract

We address the issue of judging the significance of rare events as it typically arises in statistical natural-language processing. We first define a general approach to the problem, and we empirically compare results obtained using log-likelihood-ratios and Fisher's exact test, applied to measuring strength of bilingual word associations.

1 Introduction

Since it was first introduced to the NLP community by Dunning (1993), the G^2 log-likelihood-ratio statistic[1] has been widely used in statistical NLP as a measure of strength of association, particularly lexical associations. Nevertheless, its use remains controversial on the grounds that it may be unreliable when applied to rare events. For instance Pedersen, et al. (1996) present data showing that significance values for rare bigrams estimated with G^2 can differ substantially from the true values as computed by Fisher's exact test. Although Dunning argues that G^2 is superior to the chi-square statistic[2] X^2 for dealing with rare events, Agresti (1990, p. 246) cites studies showing "X^2 is valid with smaller sample sizes and more sparse tables than G^2," and either X^2 or G^2 can be unreliable when expected frequencies of less than 5 are involved, depending on circumstances.

The problem of rare events invariably arises whenever we deal with individual words because of the Zipfian phenomenon that, typically, no matter how large a corpus one has, most of the distinct words in it will occur only a small number of times. For example, in 500,000 English sentences sampled from the Canadian Hansards data supplied for the bilingual word alignment workshop held at HLT-NAACL 2003 (Mihalcea and Pedersen, 2003), there are 52,921 distinct word types, of which 60.5% occur five or fewer times, and 32.8% occur only once.

The G^2 statistic has been most often used in NLP as a measure of the strength of association between words, but when we consider pairs of words, the sparse data problem becomes even worse. If we look at the 500,000 French sentences corresponding to the English sentences described above, we find 19,460,068 English-French word pairs that occur in aligned sentences more often than would be expected by chance, given their monolingual frequencies. Of these, 87.9% occur together five or fewer times, and 62.4% occur together only once. Moreover, if we look at the expected number of occurrences of these word pairs (which is the criteria used for determining the applicability of the X^2 or G^2 significance tests), we find that 93.2% would be expected by chance to have fewer than five occurrences. Pedersen et al. (1996) report similar proportions for monolingual bigrams in the ACL/DCI Wall Street Journal corpus. Any statistical measure that is unreliable for expected frequencies of less than 5 would be totally unusable with such data.

2 How to Estimate Significance for Rare Events

A wide variety of statistics have been used to measure strength of word association. In one paper alone (Inkpen and Hirst, 2002), pointwise mutual information, the Dice coefficient, X^2, G^2, and Fisher's exact test statistic were all computed and combined to aid in learning collocations. Despite the fact that many of these statistics arise from significance testing, the usual practice in applying them in NLP is to choose a threshold heuristically for the value of the statistic being used and discard all the pairs below the threshold. Indeed, Inkpen and Hirst say (p. 70) "there is no principled way of choosing these thresholds."

This may seem an odd statement about the measures that arise directly from significance testing,

* From *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, pp. 333–340.

[1] Dunning did not use the name G^2, but this appears to be its preferred name among statisticians (e.g., Agresti, 1990).

[2] Following Agresti, we use X^2 to denote the test statistic and χ^2 to denote the distribution it approximates.

but it is clear that if standard statistical tests are used naively, the results make no sense in these applications. One might suppose that this is merely the result of the statistics in question not being applicable to the rare events that predominate in NLP, but it is easy to show this is not so.

2.1 When is Something Seen Only Once Signficant?

Consider the case of two words that each occur only once in a corpus, but happen to co-occur. Conventional wisdom strongly advises suspicion of any event that occurs only once, yet it is easy to see that applying standard statistical methods to this case will tend to suggest that it is highly significant, without using any questionable approximations at all.

The question that significance tests for association, such as X^2, G^2, and Fisher's exact test, are designed to answer is, given the sample size and the marginal frequencies of the two items in question, what is the probability (or p-value) of seeing by chance as many or more joint occurrences as were observed? In the case of a joint occurrence of two words that each occur only once, this is trivial to calculate. For instance, suppose an English word and a French word each occur only once in our corpus of 500,000 aligned sentence pairs of Hansard data, but they happen to occur together. What is the probability that this joint occurrence happened by chance alone? We can suppose that the English word occurs in an arbitrary sentence pair. The probability that the French word, purely by chance, would occur in the same sentence pair is clearly 1 in 500,000 or 0.000002. Since it is impossible to have more than one joint occurrence of two words that each have only a single occurrence, 0.000002 is the exact p-value for the question we have asked.

Clearly, however, one cannot assume that the association of these two words is 0.999998 certain on this basis alone. The problem is that there are so many possible singleton-singleton pairs, it is very likely that some of them will occur jointly, purely by chance. This, too, is easy to calculate. In our 500,000 sentence pairs there are 17,379 English singletons and 22,512 French singletons; so there are 391,236,048 possible singleton-singleton pairs. For each pair, the probability of having a joint occurrence by chance is 0.000002, so the expected number of chance joint occurrences of singleton-singleton pairs is $391,236,048 \times 0.000002$, or approximately 782.5.

The question of whether a singleton-singleton pair is signficant or not then turns on how many singleton-singleton pairs we observe. If we see only about 800, then they are not signficant, because that is just about the number we would expect to see by chance. In our corpus, however, we see far more than that: 19,312. Thus our best estimate of the proportion of the singleton-singleton pairs that are due to chance is $782.5/19312 = 0.0405$, which we can think of as the "expected noise" in the singleton-singleton pairs. Looked at another way, we can estimate that at least 95.9% of the observed singleton-singleton pairs are not due to chance, which we can think of as "expected precision".[3] So, we conclude that, for this data, seeing two singletons together is significant at the 0.05 level, but this is more than five orders of magnitude less significant than naive use of standard p-values would suggest.

2.2 Generalizing the Method

In the previous section, we used the p-value for the observed joint frequency given the marginal frequencies and sample size as a our base statistical measure. We used this in an indirect way, however, that we could apply to any other measure of association. For example, for a joint occurrence of two singletons in 500,000 samples, G^2 is approximately 28.24. Therefore, if we wanted to use G^2 as our measure of association, we could compare the number of word pairs expected by chance to have a G^2 score greater than or equal to 28.24 with the number of word pairs observed to have a G^2 score greater than or equal to 28.24, and compute expected noise and precision just as we did with p-values. In principle, we can do the same for any measure of association. The worst that can happen is that if the measure of association is not a good one (i.e., if it assigns values randomly), the expected precision will not be very good no matter how high we set the threshold.

This means that we can, if we wish, use two different statistics to estimate expected noise and precision, one as a measure of association and one to estimate the number of word pairs expected by chance to have a given level or higher of the association measure. In our experiments, we will use a likelihood-ratio-based score as the measure of association, and contrast the results obtained using either a likelihood-ratio-based test or Fisher's exact

[3] Using the binomial distribution we can calculate that there is a 0.99 probability that there are no more than 848 singleton-singleton pairs by chance, and hence that there is a 0.99 probability that at least 95.6% of the observed singleton-singleton pairs are not due to chance. Since this differs hardly at all from the expected precision, and there is no a priori reason to think overestimating precision is any worse than underestimating it, we will use expected values of noise and precision as our primary metrics in the rest of the paper.

test to estimate expectations.

Computing the expected number of pairs with a given association score or higher, for a large collection of word pairs having a wide range of marginal frequencies, turns out to be somewhat tricky. We must first compute the p-value for an association score and then multiply the p-value by the appropriate number of word pairs. But if the association score itself does not correlate exactly with p-value, the relationship between association score and p-value will vary with each combination of marginal frequencies.[4] Furthermore, even for a single combination of marginal frequencies, there is in general no way to go directly from an association score to the corresponding p-value. Finally, until we have computed all the expected frequencies and observed frequencies of interest, we don't know which association score is going to correspond to a desired level of expected precision.

These complications can be accomodated as follows: First compute the distinct marginal frequencies of the words that occur in the corpus (separately for English and French), and how many distinct words there are for each marginal frequency. Next, choose a set of association score thresholds that we would like to know the expected precisions for.

Accumulate the expected pair counts for each threshold by iterating through all possible combinations of observed marginals. For each combination, compute the association score for each possible joint count (given the marginals and the sample size), starting from the smallest one greater than the expected joint count $C(x)C(y)/N$ (where $C(x)$ and $C(y)$ are the marginals and N is the sample size). Whenever the first association score greater than or equal to one of the thresholds is encountered, compute the associated p-value, multiply it by the number of possible word pairs corresponding to the combination of marginals (to obtain the expected number of word pairs with the given marginals having that association score or higher), and add the result to the accumluators for all the thresholds that have just been passed. Stop incrementing the possible joint frequency when either the smaller of the two marginals is reached or the highest association threshold is passed. (See Figure 1 for the details.)

At this point, we have computed the number of word pairs that would be expected by chance alone

```
for each observed C(x) {
  for each observed C(y) {
    possible_pairs =
      |values of x with frequency C(x)| ×
      |values of y with frequency C(y)| ;
    C_0(x,y) = int(C(x)C(y)/N) + 1 ;
    i = 1 ;
    loop: for each C(x,y) such that
      C_0(x,y) ≤ C(x,y) ≤ min(C(x), C(y)) {
      score = assoc(C(x,y), C(x), C(y), N) ;
      if (score ≥ threshold[i]) {
        prob = p-value(C(x,y), C(x), C(y), N) ;
        expected_pairs = prob × possible_pairs ;
        while (score ≥ threshold[i]) {
          expected_count[i] += expected_pairs ;
          if (i < number of thresholds) {
            i++ ;
          }
          else {
            exit loop ;
          }
        }
      }
    }
  }
}
```

Figure 1: Algorithm for Expected Counts.

to have an association score equal to or greater than each of our thresholds. Next we compute the number of word pairs observed to have an association score equal to or greater than each of the thresholds. The expected noise for each threshold is just the ratio of the expected number of word pairs for the threshold to the observed number of word pairs for the threshold, and the expected precision is 1 minus the expected noise.

What hidden assumptions have we made that could call these estimates into question? First, there might not be enough data for the estimates of expected and observed frequencies to be reliable. This should seldom be a problem in statistical NLP. For our 500,000 sentence pair corpus, the cumulative number of observed word pairs is in the tens of thousands for for any association score for which the estimated noise level approaches or exceeds 1%, which yields confidence bounds that should be more than adequate for most purposes (see footnote 3).

A more subtle issue is that our method may overestimate the expected pair counts, resulting in excessively conservative estimates of precision. Our estimate of the number of pairs seen by chance for a

[4] We assume that for fixed marginals and sample size, and joint frequencies higher than the expected joint frequency, the association score will increase monotonically with the joint frequency. It is hard to see how any function without this property could be considered a measure of association (unless it decreases monotonically as joint frequency increases).

$$2\log\left[\frac{p(y|x)^{C(x,y)}\cdot p(y|\neg x)^{C(\neg x,y)}\cdot p(\neg y|x)^{C(x,\neg y)}\cdot p(\neg y|\neg x)^{C(\neg x,\neg y)}}{p(y)^{C(y)}\cdot p(\neg y)^{C(\neg y)}}\right] \tag{1}$$

$$2\log\left[\frac{p(y|x)^{C(x,y)}\cdot p(y|\neg x)^{C(\neg x,y)}\cdot p(\neg y|x)^{C(x,\neg y)}\cdot p(\neg y|\neg x)^{C(\neg x,\neg y)}}{p(y)^{C(x,y)}\cdot p(y)^{C(\neg x,y)}\cdot p(\neg y)^{C(x,\neg y)}\cdot p(\neg y)^{C(\neg x,\neg y)}}\right] \tag{2}$$

$$2\log\prod_{x?\in\{x,\neg x\}}\prod_{y?\in\{y,\neg y\}}\left(\frac{p(y?|x?)}{p(y?)}\right)^{C(x?,y?)} \tag{3}$$

$$2\left[\sum_{x?\in\{x,\neg x\}}\sum_{y?\in\{y,\neg y\}}C(x?,y?)\log\frac{p(y?|x?)}{p(y?)}\right] \tag{4}$$

$$2N\left[\sum_{x?\in\{x,\neg x\}}\sum_{y?\in\{y,\neg y\}}p(x?,y?)\log\frac{p(x?,y?)}{p(x?)p(y?)}\right] \tag{5}$$

Figure 2: Alternative Formulas for G^2.

particular value of the association measure is based on considering all possible pairs as nonassociated, which is a valid approximation only if the number of pairs having a significant positive or negative association is very small compared to the total number of possible pairs.

For the corpus used in this paper, this seems unlikely to be a problem. The corpus contains 52,921 distinct English words and 66,406 distinct French words, for a total of 3,514,271,926 possible word pairs. Of these only 19,460,068 have more than the expected number of joint occurrences. Since most word pairs have no joint occurrences and far less than 1 expected occurrence, it is difficult to get a handle on how many of these unseen pairs might be negatively associated. Since we are measuring association on the sentence level, however, it seems reasonable to expect fewer word pairs to have a significant negative association than a positive association, so 40,000,000 seems likely to be a upper bound on how many word pairs are significantly nonindependent. This, however, is only about 1% of the total number of possible word pairs, so adjusting for the pairs that might be significantly related would not make an appreciable difference in our estimates of expected noise. In applications where the significantly nonindependent pairs do make up a substantial proportion of the total possible pairs, an adjustment should be made to avoid overly conservative estimates of precision.

3 Understanding G^2

Dunning (1993) gives the formula for the statistic we are calling G^2 in a form that is very compact, but not necessarily the most illuminating:

$$2\,[\log L(p_1,k_1,n_1)+\log L(p_2,k_2,n_2)-\log L(p,k_1,n_1)-\log L(p,k_2,n_2)\,],$$

where

$$L(p,k,n)=p^k(1-p)^{n-k}.$$

The interpretation of the statistic becomes clearer if we re-express it in terms of frequencies and probabilities as they naturally arise in association problems, as shown in a number of alternative formulations in Figure 2. In these formulas, x and y represent two words for which we wish to estimate the strength of association. $C(y)$ and $C(\neg y)$ are the observed frequencies of y occurring or not occurring in the corpus; $C(x,y),\ldots,C(\neg x,\neg y)$ are the joint frequencies of the different possible combinations of x and y occuring and not occuring; and $p(y),p(\neg y),p(y|x),\ldots,p(\neg y|\neg x)$ are the maximum likelihood estimates of the corresponding marginal and conditional probabilities.

Formula 1 expresses G^2 as twice the logarithm of a ratio of two estimates of the probability of a sequence of observations of whether y occurs; one estimate being conditioned on whether x occurs, and the other not. The estimate in the numerator is conditioned on whether x occurs, so the numerator is a product of four factors, one for each possible combination of x occuring and y occuring. The overall probability of the sequence is the product of each conditional probability of the occurrence or nonoccurrence of y conditioned on the occurrence or nonoccurrence of x, to the power of the number of times the corresponding combination occurs in the sequence of observations. The denominator is

an estimate of the probability of the same sequence, based only on the marginal probability of y. Hence the denominator is simply the product of the probabilty of y occuring, to the power of the number of times y occurs, and the probabilty of y not occuring, to the power of the number of times y fails to occur.[5]

The rest of Figure 2 consists of a sequence of minor algebraic transformations that yield other equivalent formulas. In Formula 2, we simply factor the denominator into four factors corresponding to the same combinations of occurrence and nonoccurrence of x and y as in the numerator. Then, by introducing $x?$ and $y?$ as variables ranging over the events of x and y occurring or not occurring, we can re-express the ratio as a doubly nested product as shown in Formula 3. By distributing the log operation over all the products and exponentiations, we come to Formula 4. Noting that $C(x?, y?) = N \cdot p(x?, y?)$ (where N is the sample size), and $p(y?|x?)/p(y?)$ times $p(x?)/p(x?)$ is $p(x?, y?)/p(x?)p(y?)$, we arrive at Formula 5. This can be immediately recognized as $2N$ times the formula for the (average) mutual information of two random variables,[6] using the maximum likelihood estimates for the probabilities involved.

The near equivalence of G^2 and mutual information is important for at least two reasons. First, it gives us motivation for using G^2 as a measure of word association that is independent of whether it usable for determining significance. Mutual information can be viewed as a measure of the information gained about whether one word will occur by knowing whether the other word occurs. A priori, this is at least as plausible a measure of strength of association as is the degree to which we should be surprised by the joint frequency of the two words. Thus even if G^2 turns out to be bad for estimating significance, it does not follow that it is therefore a bad measure of strength of association.

The second benefit of understanding the relation between G^2 and mutual information is that it answers the question of how to compare G^2 scores as measures of strength of association, when they are obtained from corpora of different sizes. Formula 5 makes it clear that G^2 scores will increase linearly with the size of the corpus, assuming the relevant marginal and conditional probabilities remain the same. The mutual information score is independent of corpus size under the same conditions, and thus offers a plausible measure to be used across corpora of varying sizes.

4 Computing Fisher's Exact Test

In Section 2 we developed a general method of estimating significance for virtually any measure of association, given a way to estimate the expected number of pairs of items having a specified degree of association or better, conditioned on the marginal frequencies of the items composing the pair and the sample size. We noted that for some plausible measures of association, the association metric itself can be used to estimate the p-values needed to compute the expected counts. G^2 is one such measure, but it is questionable whether it is usable for computing p-values on the kind of data typical of NLP applications. We will attempt to answer this question empirically, at least with respect to bilingual word association, by comparing p-values and expected noise estimates derived from G^2 to those derived from a gold standard, Fisher's exact test.

In this test, the hypergeometric probability distribution is used to compute what the exact probability of a particular joint frequency would be if there were no association between the events in question, given the marginal frequencies and the sample size. The only assumption made is that all trials are independent. The formula for this probability in our setting is:

$$\frac{C(x)!\,C(\neg x)!\,C(y)!\,C(\neg y)!}{N!\,C(x,y)!\,C(\neg x,y)!\,C(x,\neg y)!\,C(\neg x,\neg y)!}$$

The p-value for a given joint frequency is obtained by summing the hypergeometric probability for that joint frequency and every more extreme joint frequency consistent with the marginal frequencies. In our case "more extreme" means larger, since we are only interested in positive degrees of association.[7] Because it involves computing factorials of potentially large numbers and summing

[5] The interpretation of G^2 in terms of a likelihood ratio for a particular sequence of observations omits the binomial coefficients that complicate the usual derivation in terms of all possible sequences having the observed joint and marginal frequencies. Since all such sequences have the same probability for any given probability distributions, and the same number of possible sequences are involved in both the numerator and denominator, the binomial coefficients cancel out, yielding the same likelihood ratio as for a single sequence.

[6] After discovering this derivation, we learned that it is, in fact, an old result (Attneave, 1959), but it seems to be almost unknown in statistical NLP. The only reference to it we have been able to find in the statistical NLP "literature" is a comment in Pederson's publically distributed Perl module for computing mutual information (http://search.cpan.org/src/TPEDERSE/Text-NSP-0.69/ Measures/tmi.pm). It can also be seen as a special case of a more general result presented by Cover and Thomas (1991, p. 307, 12.187–12.192), but otherwise we have not found it in any contemporary textbook.

[7] The null hypothesis that we wish to disprove is that a

over many possible joint frequencies, this test has traditionally been considered feasible only for relatively small sample sizes. However, a number optimizations enable efficient estimation of p-values by Fisher's exact test for sample sizes up to at least 10^{11} on current ordinary desktop computers, where the limiting factor is the precision of 64-bit floating point arithmetic rather than computation time.

Some keys to efficient computation of Fisher's exact test are:

- The logarithms of factorials of large numbers can be efficiently computed by highly accurate numerical approximations of the gamma function (Press et al., 1992, Chapter 6.1), based on the identity $n! = \Gamma(n+1)$.

- The following well-known recurrence relation for the hypergeometric distribution:

$$P_k = \frac{C_{k-1}(\neg x, y)\, C_{k-1}(x, \neg y)}{C_k(x, y)\, C_k(\neg x, \neg y)} P_{k-1}$$

  makes it easy to calculate probabilities for a sequence of consecutive joint frequencies, once the first one is obtained. (The subscript k indicates parameters associated with the kth joint frequency in the sequence.)

- The highest possible joint frequency will be the smaller of the two marginal frequencies, so if one of the marginals is small, few terms need to be summed.

- If we iterate from less extreme joint frequencies to more extreme joint frequencies, each probability in the summation will be smaller than the one before. If both the marginals are large, the summation will often converge to a constant value, given limited arithmetic precision, long before the smaller marginal is reached, at which point we can stop the summation.

By taking advantage of these observations, plus a few other optimizations specific to our application, we are able to estimate the necessary expected joint frequencies for our 500,000 sentence pair corpus in 66.7 minutes using Fisher's exact test, compared to 57.4 minutes using an approximate estimate based on likelihood ratios, a time penalty of only 16% for using the exact method.

pair of words is either negatively associated or not associated; hence, a one-sided test is appropriate.

5 Estimating P-Values with Log-Likelihood-Ratios

The usual way of estimating p-values from log-likelihood-ratios is to rely on the fact that the p-values for G^2 asymtotically approach the well-understood χ^2 distribution, as the sample size increases. This is subject to the various caveats and conditions that we discussed in Section 1, however. Since we have the ability to compute all of the exact p-values for our corpus, we do not need to rely on the χ^2 approximation to test whether we can use log-likelihood-ratios to estimate p-values. We can empirically measure whether there is any consistent relationship between log-likelihood-ratios and p-values, and if so, use it empirically to estimate p-values from log-likelihood-ratios without resorting to the χ^2 approximation. For all we know at this point, it may be possible to empirically predict p-values from G^2 under conditions where the correspondence with χ^2 breaks down.

This means we can drop the heretofore mysterious factor of 2 that has appeared in all the formulas for G^2, since this factor seems to have been introduced just to be able to read p-values directly from standard tables for the χ^2 distribution. To make it clear what we are doing, from this point on we will use a statistic we will call LLR which we define to be $G^2/2$.

To look for a relationship between LLR and p-values as computed by Fisher's exact test, we first computed both statistics for a various combinations of joint frequency, marginal frequencies, and sample sizes. Exploratory data analysis suggested a near-linear relationship between LLR scores and the negative of the natural logarithm of the p-values. To make sure the apparent relationship held for a real dataset, we computed the LLR scores and negative log p-values for all 19,460,068 English-French word pairs in our corpus with more joint occurrences than expected by chance, and carried out a least-squares linear regression, treating LLR score as the independent variable and negative log p-value as the dependent variable, to see how well we can predict p-values from LLR scores. The results are as follows:

slope:	1.00025
intercept:	1.15226
Pearson's r^2:	0.999986
standard deviation:	0.552225

With an r^2 value that rounds off to five nines, LLR score proves to be a very good predictor of the negative log p-values over the range of values considered. Moreover, with a slope of very close to

1, the LLR score and negative log p-values are not merely correlated, they are virtually the same except for the small delta represented by the intercept. In other words,

$$\text{p-value} \approx e^{-(LLR+1.15)}$$

would seem to be not too bad an approximation.

The standard deviation of 0.55, however, is at least slightly worrying. As a range of differences in the logarithms of the predicted and actual p-values, it corresponds to a range of ratios between the predicted and actual p-values from about 0.57 to 1.7.

6 Estimating Noise in Bilingual Word Association

For our final experiment, we estimated the noise in the bilingual word associations in our data by the method of Section 2, using both Fisher's exact test and LLR scores via our regression equation to estimate expected pair counts. In both cases, we use LLR scores as the measure of association. We computed the cumulative expected noise for every integral value of the LLR score from 1 through 20.

To try to determine the best results we could get by using LLR scores to estimate expected noise in the region where we would be likely to set a cut-off threshold, we recomputed the least-squares fit of LLR and negative log p-value, using only data with LLR scores between 5 and 15.[8] We obtained the following values for the parameters of the regression equation from the re-estimation:

slope:	1.04179
intercept:	0.793324

Note that the re-estimated value of the intercept makes it closer to the theoretical value, which is $-\log(0.5) \approx 0.69$, since independence corresponds to an LLR score of 0 and a p-value of 0.5.

The results of these experiments are summarized in Table 1. The first column shows the potential LLR association score cut-offs, the second column is the expected noise for each cut-off estimated by Fisher's exact test, the third column gives the noise estimates derived from p-values estimated from LLR scores, and the fourth column shows the ratio between the two noise estimates. If we look at the noise estimates based on our gold standard, Fisher's exact test, we see that the noise level is below 1% above an LLR score of 11, and rises rapidly below that. This confirms previous annecdotal experience that an LLR score above 10 seems to be a reliable indicator of a significant association.

Cut-Off	Fisher Noise Est	LLR Noise Est	Ratio
1	0.624	0.792	1.27
2	0.516	0.653	1.27
3	0.423	0.384	0.91
4	0.337	0.274	0.81
5	0.256	0.183	0.71
6	0.181	0.114	0.63
7	0.119	0.0650	0.55
8	0.0713	0.0338	0.47
9	0.0394	0.0159	0.40
10	0.0205	0.00695	0.34
11	0.00946	0.00260	0.27
12	0.00432	0.000961	0.22
13	0.00136	0.000221	0.16
14	0.00137	0.000166	0.12
15	3.52e-005	2.00e-005	0.57
16	1.56e-005	8.02e-006	0.51
17	6.82e-006	3.19e-006	0.47
18	2.94e-006	1.24e-006	0.42
19	1.24e-006	4.65e-007	0.38
20	5.16e-007	1.72e-007	0.33

Table 1: Word Association Noise Estimates.

The comparison between the two noise estimates indicates that the LLR score underestimates the amount of noise except at very high noise levels. It is worst when the LLR score cut-off equals 14, which happens to be just below the LLR score (14.122) for singleton-singleton pairs. Since, for a given sample size, singleton-singleton pairs have the lowest possible expected joint count, this is probably the effect of known problems with estimating p-values from likelihood ratios when expected counts are very small.

7 Conclusions

When we use Fisher's exact test to estimate p-values, our new method for estimating noise for collections of rare events seems to give results that are quite consistent with our previous annecdotal experience in using LLR scores as a measure of word association. Using likelihood ratios to estimate p-values introduces a substantial amount of error, but not the orders-of-magnitude error that Dunning (1993) demonstrated for estimates that rely on the assumption of a normal distribution. However, since we have also shown that Fisher's exact test can be applied to this type of problem without a major

[8] This constitutes training on the test data for the sake of obtaining an upper bound on what could be achieved using LLR scores. Should we conclude that LLR scores look promising for this use, one would want to re-run the test training the regression parameters on held-out data.

computational penalty, there seems to be no reason to compromise in this regard.

8 Acknowledgements

Thanks to Ken Church, Joshua Goodman, David Heckerman, Mark Johnson, Chris Meek, Ted Pedersen, and Chris Quirk for many valuable discussions of the issues raised in this paper. Thanks especially to Joshua Goodman for pointing out the existence of fast numerical approximations for the factorial function, and to Mark Johnson for helping to track down previous results on the relationship between log-likelihood-ratios and mutual information.

References

Alan Agresti. 1990. *Categorical Data Analysis*. John Wiley & Sons, New York, New York.

Fred Attneave. 1959. *Applications of Information Theory to Psychology*. Holt, Rinehart and Winston, New York, New York.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. John Wiley & Sons, New York, New York.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Diana Z. Inkpen and Graeme Hirst. 2002. Acquiring collocations for lexical choice between near-synonyms. In *Unsupervised Lexical Acquisition: Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pp. 67–76, Philadelphia, Pennsylvania.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop, Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pp. 1–6, Edmonton, Alberta.

Ted Pedersen, Mehmet Kayaalp, and Rebecca Bruce. 1996. Significant Lexical Relationships. In *Proceedings of the 13th National Conference on Artificial Intelligence*, Portland, Oregon.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipies in C: The Art of Scientific Computing, Second Edition*. Cambridge University Press, Cambridge, England.