Trading Recall for Precision with Confidence Sets

Mark Johnson*
Brown University

Many statistical applications in computational linguistics are plagued by sparse data. Exact statistics can be used to avoid the inaccuracies of methods that rely on large sample approximations, but point estimators (even using exact statistics) suffer from high variance when used with extremely small sample sizes. While the performance of any statistical method is limited by the lack of information in a small sample, confidence set estimators provide a systematic way of trading recall for precision, e.g., increasing the likelihood that an association actually exists, at the cost of potentially ignoring many real associations.

1 Introduction

Statistical estimation plays an important role in modern computational linguistics. Because many linguistic phenomena obey a Zipf's Law distribution where most types occur with low frequency, we often face estimation or inference problems with data drawn from very small samples. These small sample sizes introduce two distinct problems:

1. Estimates from small samples have *high variance*, i.e., they are likely to be unreliable. If the experiment were repeated again with new data very different estimates might well be obtained.

2. Familiar statistical methods based on *large sample approximations are likely to be inaccurate* on small samples. Informally, most statistical methods become more accurate as $n \to \infty$, whereas $n = 1$ is perhaps the most common case in computational linguistics.

* This work was supported by NSF awards 9720368, 9870676 and 9812169. I would like to thank Eugene Charniak, Brian Roark and the other members of the Brown Laboratory for Linguistic Information Processing (BLLIP), as well as Chris Manning, for helpful comments.

One standard method of dealing with these two problems is to fix a minimum sample size (say, 5 occurences) and simply ignore any events that occur less frequently. This method often works well in practice, but in some applications it is difficult to tell just how to set this minimum size, and such an approach often results in discarding the bulk of the data.

The second difficulty can be addressed by using statistical methods that are accurate for small sample sizes. Recently various *exact statistics* have been developed which are accurate for all sample sizes. These methods are often computationally intensive, usually involving either randomization and resampling or explicit enumeration, but in many applications they are quite feasible. Exact statistics are used in both the examples discussed below.

Exact statistics do not address high variance of point estimates from small samples. This difficulty is fundamental: small samples simply do not yield much information. While this limitation cannot be avoided, sometimes we can minimize its effects by reformulating our problem so it requires less information. Geman, Bienenstock, and Doursat (1992) explain this in terms of what they call the *bias/variance trade-off*. The variance of an estimator reflects the amount of variation in the estimated model from run to run; all else equal we prefer estimators with lower variance because the estimates they produce are likely to contain less "noise". The bias of an estimator is the difference between the expected value of estimated models and the true model (conceptually, averaged over many runs); informally, an unbiased estimator is one whose expected output is the correct model. Geman et. al. point out that bias and variance are inextricably related: reducing bias in general increases variance. Informally, with training data of fixed size, a simpler, biased model may perform better than a more complex and potentially more accurate model because the simpler model can be estimated more accurately from the available data; i.e., by trading variance for bias, one may obtain better overall perfor-

mance.

Since the variance of an estimated model is higher when the amount of training data is smaller, trading variance for bias is often desirable when faced with sparse data. Sometimes there are obvious ways to simplify a model in order to reduce variance at the expense of introducing bias; e.g., in a neural network reducing the number of hidden units reduces the number of parameters to be estimated (and usually the overall model variance), but may result in a network that cannot fit the data as well (Geman, Bienenstock, and Doursat, 1992).

However, it is not always obvious how to trade off variance for bias in a given type of model, especially when the model involves only a single parameter. This kind of model is common in computational linguistics—the models of verb-object selection and word pair association discussed below involve only a single parameter—so Geman et. al.'s strategy of reducing the number of parameters cannot be applied here.

Instead, this paper shows how confidence set estimators provide a general way of implementing a kind of bias/variance trade-off which can be applied even to single parameter models. In contrast to the more well-known point estimators (which select a particular model given training data), a *confidence set estimator* identifies a set of possible models. Confidence set estimators provide a systematic way of *trading recall for precision* in the computational linguistic applications involving association detection discussed below. This trade-off is controlled by the user-specified confidence-level parameter α. As α is lowered fewer associations are found, but it is more likely that the associations are real, rather than due to statistical fluctuation. The details of how such confidence sets are used to trade recall for precision is discussed in section 2.

The rest of this introduction briefly reviews some of the previous work on measures of association especially relevant to the confidence set approach. I want to make clear that this paper is not introducing a new measure of association; rather, it introduces a

method of trading recall for precision which can in principle be used with any measure of association, and should be relevant to other applications besides estimating association.

There is no single standard measure of "degree of association" in the statistical literature (Press et al., 1992), and many different measures of association have been proposed in computational linguistics. It seems plausible that a measure that works well in one application may fare poorly in another application, so there may be no single "best" measure of association.

It is important to distinguish the quantity we want to measure from estimators of that quantity, which map training data to estimates of that quantity. For example, the mutual information of a pair of binomial distributions is a function of their "success probabilities". If we have samples drawn from these two distributions we can estimate the distributions' success probability parameters, and from these estimate the mutual information of the pair of distributions. There are two quite distinct reasons why the value produced by such an estimator might not be a good measure of association:

1. the estimate of the quantity is not close to the true value of the quantity (because of small sample fluctuation, etc.), and

2. the (true) quantity itself is simply not a good measure of association, i.e., even if it could be estimated accurately, the quantity simply is not closely correlated with our (intuitive) concept of association.

To the extent to which this distinction is made, most of the computational linguistics work on measures of association addresses (2) (correctly, in my view). This paper, on the other hand, focusses on (1). In principle the confidence set approach is compatible with a wide variety of the existing measures of association, and the particular measures used below were chosen for their simplicity.

Identifying associations is especially relevant as an application for the confidence set method because several well-known approaches to identifying associations rely on confidence sets at least conceptually. These approaches define the degree of association of a pair of elements (e.g., words) to be the value of a statistic from a standard test of independence, such as the t-test or the Chi-squared test of association in binary (2×2) contingency tables (Church et al., 1991). Such an approach defines association to be related to how "surprising" the frequency of occurence of the associated pair is compared to the frequency of its elements, or alternatively, how unlikely it is that the same distribution generates both the observations of the pair and the individual elements: the idea is that strongly associated pairs will be unlikely to be generated from the same distributions as the elements that make them up. As Dunning (1993) points out, likelihood ratio tests are often acceptably accurate for relatively small sample sizes and are easy to calculate. Pedersen (1996) observes that Fisher's exact test of independence (which is closely related to the conditional odds ratio estimator discussed below) is accurate for samples of any size.

Undoubtedly the best choice for a measure of association depends on the intended application, and there is no reason to suppose that a single measure will be optimal for all applications. However, for many applications statistics derived from tests of independence are probably not the most appropriate measure of association. These statistics measure how surprising the observations of a pair are given the null hypothesis of no association. Informally, the same level of surprise can be obtained either by having a moderate number of samples from very different distributions, or else by having an enormous number of samples from only slightly different distributions: in both cases we may be equally confident that the null hypothesis is false. In practice, this means that the highest scores for these tests are obtained from very frequently occuring words, even though intuitively these words may not be closely associated. A quick glance through

Table 2 shows that the likelihood ratio test tends to rank highly extremely frequent pairs. The corresponding confidence set approach seeks a lower bound on a measure of association (the odds ratio) and it turns out that we can be reasonably confident that a strong association exists given a much more modest amount of information.

A second, technical, problem with adopting standard tests of independence is that these tests are two-sided; e.g., a bigram scores highly if it occurs either more *or less* frequently than would be expected given the distribution of each word individually. For example, the bigram *the the* scores highly on a likelihood ratio test precisely because it appears much less frequently than the frequency of the word *the* would lead us to expect. (This problem could be avoided by using one-sided rather than two-sided tests).

2 Confidence set estimation

This section explains how confidence set estimators can be used to increase the precision in identifying associations at the expense of reducing recall, or equivalently, introducing a bias toward no association. This use of confidence set estimators is quite general, and is applicable to other quantities besides measures of association.

A confidence set estimator maps a training data sample to a set S_α of possible parameter values or models. One method of identifying a confidence set S_α is as follows. A model m is in S_α iff m generates the observed sample datum x as well as all other less likely possible sample data (w.r.t. m) with (cumulative) probability greater than or equal to α. The confidence level α is a user-specified parameter. In more detail, let x be the observed sample datum, $\Pr_m(x)$ be the probability of x with respect to model m, D be the set of possible sample data (e.g., all possible observations), M be the set of possible models (e.g., all possible values of model parameters), and α be a user-specified confidence level. Define $D_{m,x}$ to be the set of possible sample data that are not more

likely than the observed sample data x under m, i.e.,

$$D_{m,x} \;=\; \{x' \in D | \Pr_m(x') \leq \Pr_m(x)\}$$

Then the confidence set S_α contains all models m such that $\Pr_m(D_{m,x}) \geq \alpha$, i.e.,

$$S_\alpha \;=\; \left\{ m \in M | \int_{D_{m,x}} \Pr_m(x').dx' \geq \alpha \right\}.$$

The quantity v we want to estimate is typically a function of the model or model parameters m, i.e., $v = v(m)$. Given a confidence set S_α, we use the minimum value that v takes on S_α as our estimator $\hat{v}_-$ of v. That is,

$$\hat{v}_- = \operatorname*{argmin}_{m \in S_\alpha} v(m).$$

An example may make this clearer. Suppose we want to identify pairs (v, n) of verbs v and nouns n such that n is likely to appear as the head of the direct object of v, e.g., $v =$ *evade*, $n =$ *taxes*. While many measures of association could plausibly be used, for simplicity here we use the parameter $\theta = \Pr(N{=}n|V{=}v)$, the probability of seeing noun n as the head of the direct object of verb v, as a measure of association. As explained above, the confidence set approach is not restricted to using this measure of association; nevertheless it is surprising how well such a simple quantity performs.

Given a corpus of verb-direct object pairs, the mean $\hat{\theta} = C(v,n)/C(v)$ is a well-known point estimator for $\theta = \Pr(N{=}n|V{=}v)$, where $C(n,v)$ and $C(v)$ are the number of times the (v,n) combination and v occured in the corpus respectively. As is well-known, the mean is unreliable with extremely small sample sizes. For example, it reaches its maximum possible value, 1, on verb-noun combinations where both the pair and the verb occur exactly once in the corpus such as *convicts congressman*, even though *convicts* does not strongly select *congressman* (although some might argue that this small sample is representative).

Intuitively, we should "discount" the estimates for $\hat{\theta}$ when the sample size $C(v)$ is

small, since in those cases a large value for $\hat{\theta}$ could have arisen by small sample fluctuation, and not reflect a large value of θ. A confidence interval estimator provides a systematic way of performing this discounting. Specifically, we use the Clopper-Pearson confidence interval estimator which maps the sample data counts $C(v,n)$ and $C(v)$ to the lower bound $\hat{\theta}_-$ of a confidence interval for the "true" model parameter θ. This lower bound $\hat{\theta}_-$ is used as a "discounted" measure of association. The confidence level α is a user-specified parameter which determines how sparse data should be discounted; as $\alpha \to 0$ small samples are increasingly discounted. Calculating $\hat{\theta}_-$ is considerably more complicated than $\hat{\theta}$; it involves numerical solution of a nonlinear equation (see the appendix for downloadable software). Table 1 shows the results of such an analysis applied to counts from the U. Penn Wall Street Journal treebank (Marcus, Santorini, and Marcinkiewicz, 1993); at $\alpha = 0.1$ pairs with small sample sizes are ranked highly, while the pairs ranked most highly at $\alpha = 0.0001$ involve considerably higher larger samples.

A large value of $\hat{\theta}_-$ indicates that the true model parameter $\theta = \Pr(N{=}n|V{=}v)$ is itself likely to be large, but a small value of $\hat{\theta}_-$ could be due either to the true θ being small, or a small sample size. Thus by reducing α one effectively trades recall for precision in identifying association: the pairs identified as associated are more likely to in fact be associated, but fewer associated pairs with small samples will be detected.

This can also be viewed as an implementation of Geman et. al.'s bias-variance tradeoff in the following sense. The mean $\hat{\theta} = C(v,n)/C(v)$ is an unbiased estimator of $\theta = \Pr(N{=}n|V{=}v)$ that is optimal in a certain statistical sense; if one must estimate θ on the basis of the counts $C(v,n)$ and $C(v)$ alone in general it is not be possible to find a better estimator than the mean.[1] On the other hand, the lower bound $\hat{\theta}_-$ of the confidence

1 Note that in the verb-object selection application we do have additional information concerning how other verbs select for their objects, and it might be reasonable to assume that most verbs have similiar distributions of selected objects. This information might be incorporated into prior in a Bayesian approach.

$\alpha = 0.1$			$\alpha = 0.0001$		
k/n	$\hat{\theta}_-$	*verb object*	k/n	$\hat{\theta}_-$	*verb object*
157/220	0.670947	yield %	157/220	0.590458	yield %
4/4	0.562341	evade taxes	12/23	0.166661	controls %
4/4	0.562341	dilute earnings	13/27	0.160536	sustained damage
3/3	0.464159	bribed officials	13/27	0.160536	indicating coupon
7/10	0.448269	override veto	35/118	0.157075	owns %
8/12	0.440997	yielding %	8/12	0.156923	yielding %
4/5	0.416110	exercises option	7/10	0.143361	override veto
4/5	0.416110	dominates market	12/35	0.100922	carries warrant
12/23	0.370116	controls %	21/86	0.100765	holds %
13/27	0.344807	sustained damage	4/4	0.100000	evade taxes
13/27	0.344807	indicating coupon	4/4	0.100000	dilute earnings
4/6	0.333194	solving problems	8/17	0.099233	veto bill
4/6	0.333194	revoke license	18/70	0.097530	totaled shares
4/6	0.333194	narrows return	22/100	0.092343	earned cents
3/4	0.320461	ruining market	10/28	0.090052	compares profit
2/2	0.316228	underperform stocks	10/29	0.086480	financing order
2/2	0.316228	subpoena papers	9/27	0.074914	post loss
2/2	0.316228	strengthens links	12/46	0.074273	losing money
2/2	0.316228	spraying dispersant	6/12	0.073668	obtaining financing
2/2	0.316228	sever ties	9/28	0.071865	lower rates
2/2	0.316228	reserve right	4/5	0.067813	exercises option
2/2	0.316228	rescinding order	4/5	0.067813	dominates market
2/2	0.316228	pave way	7/18	0.067044	killing people
2/2	0.316228	overhanging market	6/13	0.066672	filling vacancy
2/2	0.316228	outlawing abortion	10/40	0.060257	lowered ratings
2/2	0.316228	inducing immunity	15/84	0.058461	signed agreement
2/2	0.316228	impede trade	13/66	0.058373	own %
2/2	0.316228	grounding airline	43/403	0.057890	reported loss
2/2	0.316228	footing bill	10/43	0.055663	represent transactions
2/2	0.316228	fatten cattle	11/53	0.053717	played role
2/2	0.316228	exonerated trading	12/62	0.053703	return calls
2/2	0.316228	corner market	21/156	0.053621	posted loss
2/2	0.316228	bucking trend	6/16	0.051947	withdrew offer
2/2	0.316228	booking revenue	6/16	0.051947	solve problems
2/2	0.316228	bashing government	4/6	0.051901	solving problems
2/2	0.316228	averting strike	4/6	0.051901	revoke license
2/2	0.316228	abandons efforts	4/6	0.051901	narrows return
8/17	0.297257	veto bill	41/423	0.051589	buy shares
6/12	0.288172	obtaining financing	19/142	0.050299	reached agreement
6/13	0.263730	filling vacancy	23/192	0.049891	filed suit
3/5	0.246636	weighs pounds	9/39	0.049676	pursue interests
3/5	0.246636	seizing assets	3/3	0.046416	bribed officials
3/5	0.246636	equal %	14/95	0.045398	changed hands
3/5	0.246636	computerizing operations	11/62	0.045397	play role
3/5	0.246636	calculating tax	6/18	0.045302	repay debt
35/118	0.241466	owns %	6/18	0.045302	owning %
4/8	0.239662	assessing damage	5/12	0.043940	stabilizing level
12/35	0.235379	carries warrant	6/19	0.042583	plays role
10/28	0.235018	compares profit	11/66	0.042475	reach agreement
7/18	0.231390	killing people	9/46	0.041531	obtain financing
10/29	0.226415	financing order	9/46	0.041531	declared dividend
5/12	0.218681	stabilizing level	10/58	0.040313	involving losses
9/27	0.212218	post loss	12/86	0.037957	holds stake
6/16	0.210413	withdrew offer	17/156	0.037923	included gain
6/16	0.210413	solve problems	8/40	0.037370	lowered rating
4/9	0.210396	project image	5/14	0.036631	export feet
4/9	0.210396	maximize value	6/22	0.036091	raises questions
4/9	0.210396	laying groundwork	4/8	0.035583	assessing damage
9/28	0.204201	lower rates	30/403	0.034782	reported earnings
3/6	0.200909	terminate contract	11/80	0.034666	provides services

Table 1
Verb and direct object pairs sorted by lower bound $\hat{\theta}_-$ of confidence interval estimates of the binomial parameter θ, at confidence levels $\alpha = 0.1$ and $\alpha = 0.0001$.

interval is in general a *biased* estimator of θ, i.e., its expected value is not θ.[2] By using $\hat{\theta}_-$ as a measure of association (rather than $\hat{\theta}$) the rate of false positives is decreased; in this sense the confidence interval estimator trades variance for bias as $\alpha \rightarrow 0$.[3]

3 Confidence interval estimators for binomial distributions

This section describes the Clopper-Pearson method for estimating confidence intervals for binomial distributions. Formally, we observe an event occuring k times in m Bernoulli trials from a binomial distribution, and we wish to estimate a confidence interval for the unknown success probability θ. For any θ besides 0 and 1 there is a non-zero likelihood of observing any number k of events in m trials, so it is not possible to find a non-trivial hard bound on θ. The best one can hope for is an interval such that if θ lies outside of this interval it is unlikely (but not impossible) for events such as the one observed to occur. Now the likelihood of observing i events in m trials is $\Pr_{m,\theta}(i) = \binom{m}{i}\theta^i(1-\theta)^{m-i}$. Given a confidence level α, we seek a lower bound $\hat{\theta}_-$ such that:

$$\sum_{i=k}^{m} \Pr_{m,\hat{\theta}_-}(i) \;=\; (1-\alpha)/2 \tag{1}$$

Thus, if $\theta \leq \hat{\theta}_-$ then the likelihood of observing k or more events in m trials is less than or equal to $(1-\alpha)/2$. Figure 1 depicts the relationship between k, α and $\hat{\theta}_-$.

A lower bound $\hat{\theta}_-$ can be calculated exactly as follows. The cumulative binomial distribution with event probability θ is related to the incomplete beta function I_θ in the following way:

$$\sum_{i=k}^{m} \binom{m}{i}\theta^i(1-\theta)^{m-i} \quad = \quad I_\theta(k, m-k+1)$$

2 However, the lower bound $\hat{\theta}_-$ is a *consistent* estimator of θ, since $\hat{\theta}_- \rightarrow \theta$ as the sample size $C(v) \rightarrow \infty$ for $0 < \alpha < 1$.

3 The variance of $\hat{\theta}_-$ may not decrease as $\alpha \rightarrow 0$.

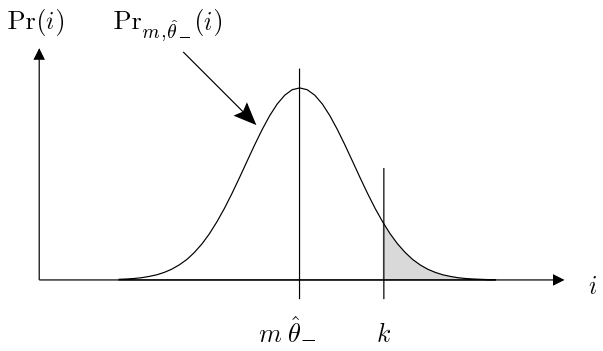

Figure 1
Calculation of an α-level confidence interval $[\hat{\theta}_-, 1]$ for the binomial parameter θ. k is the number of times the event was actually observed in m trials. The area contained in each of the shaded area above k is $(1-\alpha)/2$.

Thus the required bound is the solution of the following equation:

$$I_{\hat{\theta}_-}(k, m-k+1) \quad = \quad (1-\alpha)/2 \tag{2}$$

Equation (2) can be solved numerically using standard techniques. Press et al. (1992) explain how to efficiently calculate the incomplete beta function. Hollander and Wolfe (1999) discusses exact and asymptotic approximations for confidence intervals for the binomial parameter θ.

4 Confidence intervals for the odds ratio

The previous two sections described how to estimate confidence intervals for binomial distributions. Sometimes it is necessary to compare two different binomial distributions, and this section explains how the odds ratio can be used to do this. For example, one might say that the word bigram $w_1 w_2$ is strong associated if $\Pr(W_1 = w_1 | W_2 = w_2)$ is markedly greater than $\Pr(W_1 = w_1 | W_2 \neq w_2)$; the odds ratio quantifies such a relationship. The odds ratio is especially appropriate if one wants to estimate a confidence set, since an exact confidence set for the odds ratio can be calculated (Agresti, 1992), and it possesses several other desirable properties (Lloyd, 1999). But as noted above, this section's primary goal is to demonstrate how confidence sets can be used to trade recall for

precision rather than to argue for any particular measure of association.

Suppose we have two binomial distributions, with "success probability" parameters θ_1 and θ_2. (In the bigram association application $\theta_1 = \Pr(W_1 = w_1 | W_2 = w_2)$ and $\theta_2 = \Pr(W_1 = w_1 | W_2 \neq w_2)$). The corresponding *odds* o_1, o_2 of these distributions are:

$$o_1 = \frac{\theta_1}{1-\theta_1}, \qquad o_2 = \frac{\theta_2}{1-\theta_2}.$$

Then the *odds ratio* Δ is the ratio of the odds of the two distributions, i.e.,

$$\Delta(\theta_1, \theta_2) = \frac{\theta_1 (1-\theta_2)}{\theta_2 (1-\theta_1).} \tag{3}$$

The odds ratio varies from zero to infinity. If $\theta_1 > \theta_2$ then $\Delta > 1$, while if $\theta_1 < \theta_2$ then $\Delta < 1$. The odds ratio is *symmetric*, i.e.,

$$\Delta(\Pr(X = x | Y = y), \Pr(X = x | Y \neq y)) = \Delta(\Pr(Y = y | X = x), \Pr(Y = y | X \neq x)).$$

It seems reasonable that a measure of bigram association should be symmetric as there is no particular reason why $\Pr(w_1|w_2)$ should be more informative than $\Pr(w_2|w_1)$. (If there is in some application, then the binomial estimator discussed in the last section might be more useful).

Suppose we have two binomial samples, arranged in the rows of a 2×2 contingency table as follows:

$$\begin{matrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{matrix}$$

In our bigram application, n_{11} is the number of times the bigram $w_1 w_2$ was observed, n_{12} is the number of times w_2 was observed not preceded by w_1, n_{21} is the number of times w_1 was observed not followed by w_2, and n_{22} is the number of times a pair was observed whose first element was not w_1 and whose second element was not w_2.

The maximum likelihood estimator $\hat{\Delta}$ for the odds ratio is given by substituting the corresponding sample means for θ_1 and θ_2 in (3), i.e.;

$$\hat{\Delta} = \frac{n_{11}\, n_{22}}{n_{21}\, n_{12}}.$$

However, this estimate $\hat{\Delta}$ is unreliable when the sample size is small. Using bigram data from the same Penn WSJ treebank corpus as earlier, $\hat{\Delta}$ achieves its maximum possible value (infinity) on bigrams such as *accidentally smother* and *wrongfully imprisoning* which occur once in the corpus, and have the property that their constituent words do not occur elsewhere (i.e., $n_{21} = n_{12} = 0$). These bigrams do not seem especially highly associated, and it is likely that the high value of $\hat{\Delta}$ is due to the small sample fluctuation. Following the general strategy for trading recall for precision proposed in this paper, we use the lower bound $\hat{\Delta}_-$ of a confidence interval for the odds ratio as a "discounted" estimator for the odds ratio Δ.

There are several ways in which this can be done, and the reader is referred to Agresti (1992), Hollander and Wolfe (1999) and Lloyd (1999) for details of asymptotic approximations and exact methods for calculating confidence intervals for the odds ratio. The exact method, which is used to produce the results shown in Table 2, involves conditional inference. That is, it involves conditioning not only on the row totals $n_{11} + n_{12}$ and $n_{21} + n_{22}$, but also on the column totals $n_{11} + n_{21}$ and $n_{12} + n_{22}$. While this may be questionable in some applications (see Lloyd (1999) for discussion) it seems reasonable in the bigram association application, since it amounts to conditioning on the number of occurences of w_1 as well as w_2. The software provided in the appendix calculates the lower bound of the conditional odds ratio by numerically solving a nonlinear equation involving the sum of the hypergeometric distribution.

Table 2 contains the highest ranked bigrams according to the lower bound $\hat{\Delta}_-$ of the odds ratio at confidence levels 0.1 and 0.0001, and by the likelihood ratio test statistic popularized by Dunning (1993) as well. The likelihood ratio statistic is one of the more popular measures of bigram association in computational linguistics. It tests the null hypothesis that the two binomial samples are drawn from the same distribution, so higher scores are generally produced by extremely frequent bigrams, as explained earlier. (The

exception are bigrams such as *the the*, which are presumably highly surprising because they occur much less frequently than expected). The reader may initially be surprised to discover that the bigrams ranked highest by the "discounted" odds ratio $\hat{\Delta}_-$ consist largely of relatively ideosyncratic names, but on reflection this seems quite reasonable.

5 Conclusion

This paper's primary goal was to show how confidence set estimates provide a systematic way to trade recall for precision in a systematic way. This trade-off of precision and recall can be seen as related to the bias-variance trade-off described in Geman, Bienenstock, and Doursat (1992).

A second point made in the paper is that it may be valuable to conceptually distinguish the question of whether a particular measure or quantity actually measures what we are interested in from the question of whether a particular statistic accurately estimates that quantity. The confidence interval lower bounds reported in this paper were calculated using exact methods, which are accurate even at small sample sizes. This paper makes no claims concerning whether the quantities that were so estimated actually correspond to association.

Finally, it seems that this approach may be fruitfully extended in several ways. The most well-known exact confidence interval estimators are obtained by conditioning on the sufficient statistics for the other parameters in the model (this is why a *conditional* odds ratio estimator was used in the previous section), but this can only be done for certain measures (such as the odds ratio). It would seem that confidence sets could also be calculated using Bayesian methods, and these methods would offer greater flexibility in the kinds of measures to be estimated as well as permitting one to incorporate prior information into the model.

Likelihood ratio		$\alpha = 0.1$		$\alpha = 0.0001$	
n_{11}	$w_1\ w_2$	n_{11}	$w_1\ w_2$	n_{11}	$w_1\ w_2$
874	New York	29	Du Pont	124	Hong Kong
4969	of the	58	Navigation Mixte	58	Navigation Mixte
3930	in the	124	Hong Kong	29	Du Pont
2758	, "	20	COMMERCIAL PAPER	34	Freddie Mac
1682	, which	17	Della Femina	20	COMMERCIAL PAPER
1089	a share	34	Freddie Mac	137	Los Angeles
2202	. "	14	Guzman Cabrera	17	Della Femina
616	more than	12	Polly Peck	14	Guzman Cabrera
784	will be	17	Khmer Rouge	45	Burnham Lambert
1169	, but	10	MERRILL LYNCH	28	Las Vegas
296	Stock Exchange	10	LATE EURODOLLARS	17	Khmer Rouge
853	said it	10	INTERBANK OFFERED	24	K mart
567	has been	10	BANKERS ACCEPTANCES	12	Polly Peck
695	" says	9	TREASURY BILLS	19	Palo Alto
322	vice president	137	Los Angeles	46	Fannie Mae
474	do n't	19	Palo Alto	28	Dean Witter
242	Wall Street	28	Las Vegas	19	Puerto Rico
1180	the company	7	Zoete Wedd	10	MERRILL LYNCH
243	San Francisco	7	Pitney Bowes	10	LATE EURODOLLARS
1535	on the	7	H.F. Ahmanson	10	INTERBANK OFFERED
1062	to be	19	Puerto Rico	10	BANKERS ACCEPTANCES
9	the the	24	K mart	9	TREASURY BILLS
473	have been	28	Dean Witter	19	L.J. Hooker
378	did n't	10	READY ASSETS	243	San Francisco
278	chief executive	10	Fulton Prebon	69	Lehman Hutton
421	" We	10	LYNCH READY	10	READY ASSETS
442	this year	10	HOME LOAN	10	Fulton Prebon
1	. .	10	CALL MONEY	10	LYNCH READY
329	does n't	6	Kuala Lumpur	10	HOME LOAN
1074	, or	6	IG Metall	10	CALL MONEY
242	net income	6	Dalkon Shield	16	De Beers
181	Dow Jones	6	Bare-Faced Messiah	7	Zoete Wedd
309	year earlier	6	Aga Khan	7	Pitney Bowes
234	York Stock	46	Fannie Mae	7	H.F. Ahmanson
181	real estate	9	U S	9	U S
1645	for the	45	Burnham Lambert	10	negotiable C.D.s
223	third quarter	5	Wastewater Treatment	10	ASSETS TRUST
445	would be	5	Peat Marwick	242	Wall Street
243	wo n't	5	gon na	105	Merrill Lynch
277	at least	5	Chez Panisse	6	Kuala Lumpur
476	cents a	19	L.J. Hooker	6	IG Metall
747	the U.S.	10	negotiable C.D.s	6	Dalkon Shield
344	such as	10	ASSETS TRUST	6	Bare-Faced Messiah
137	Los Angeles	16	De Beers	6	Aga Khan
243	compared with	4	Wedding Song	26	Costa Rica
617	from $	4	PRECIOUS METALS	20	Harris Upham
260	& Co.	4	Nissho Iwai	8	Nimitz Freeway
405	" I	4	Nihon Keizai	12	Kleinwort Benson
124	Hong Kong	4	Exporting Countries	10	Marlin Fitzwater
298	last year	4	Cobb Parkway	5	Wastewater Treatment
526	company 's	4	Castle Harlan	5	Peat Marwick
150	Big Board	4	Bonwit Teller	5	gon na
314	according to	6	Rupert Murdoch	5	Chez Panisse
8	of ,	6	Hang Seng	874	New York
230	last week	8	Nimitz Freeway	20	Wells Fargo
175	composite trading	5	Violin Concerto	6	Rupert Murdoch
173	executive officer	5	Tan Sri	6	Hang Seng
193	ca n't	5	Hun Sen	13	Backer Spielvogel
607	a year	10	Marlin Fitzwater	19	Bankruptcy Code
16	to ,	3	Won Sohn	16	Beverly Hills

Table 2
Bigrams $w_1\ w_2$ sorted by likelihood ratio, and lower bound $\hat{\Delta}_-$ of a confidence interval estimate of the odds ratio, at confidence levels $\alpha = 0.1$ and $\alpha = 0.0001$. n_{11} is the number of times the bigram occured in the corpus.

Obtaining software for computing confidence intervals

Programs for computing exact lower bounds on the success probability of a binomial distribution and on the conditional odds ratio respectively can be downloaded for research purposes from http://www.cog.brown.edu/~mj.

References

Agresti, Alan. 1992. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153.

Church, Ken, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates, New Jersey, pages 115–164.

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

Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58.

Hollander, Myles and Douglas A. Wolfe. 1999. *Nonparametric statistical methods*. J. Wiley, New York.

Lloyd, Chris J. 1999. *Statistical Analysis of Categorical Data*. J. Wiley, New York.

Marcus, Michell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Pedersen, Ted. 1996. Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference (SCSUG-96)*.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, England, 2nd edition.