Significant Lexical Relationships

Ted Pedersen and **Mehmet Kayaalp** and **Rebecca Bruce**
Department of Computer Science and Engineering
Southern Methodist University
Dallas, TX 75275-0122
{pedersen,kayaalp,rbruce}@seas.smu.edu

Abstract

Statistical NLP inevitably deals with a large number of rare events. As a consequence, NLP data often violates the assumptions implicit in traditional statistical procedures such as significance testing. We describe a significance test, an exact conditional test, that is appropriate for NLP data and can be performed using freely available software. We apply this test to the study of lexical relationships and demonstrate that the results obtained using this test are both theoretically more reliable and different from the results obtained using previously applied tests.

Introduction

Statistical Natural Language Processing (NLP) seeks to make general claims about human language from an empirical study of examples of human speech or text. Empirical studies of language implicitly or explicitly define a probabilistic model for the characteristic being studied. In significance testing, the probabilistic model is a potential description of the distribution of that characteristic in the population from which the data sample was drawn. The acceptability of a potential population model is postulated as a null hypothesis and that hypothesis is tested by evaluating the *fit* of the model to the data sample (i.e., the degree to which the data sample is approximated by the model). The fit of a model is judged acceptable if the model differs from the data sample by an amount consistent with sampling variation, that is, if the value of the metric measuring the fit of the model is *statistically significant*. The assessment of statistical significance is, in a strict sense, valid only when the data has been obtained via a random sampling process. The extent to which the data are similar to a random sample determines the extent to which the results of the analysis pertain to the true population.

The importance of assessing the statistical significance of model fit is that it provides the link between the data sample and the population as a whole. This link is the confidence that a researcher can have in attributing the results of a study, based on a sample of data, to the larger population. It also allows valid comparisons between different models and work done on different data samples.

While other researchers have used significance tests to study NLP data ((Dunning 1993), (Church *et al.* 1991)) the tests used are often inappropriate for the type of data found in NLP and therefore produce erroneous results. We describe a test, an *exact conditional test*, that can be used to accurately assess the significance of a population model from a data sample comprised of both large and small counts or where many of the counts are zero. We apply the exact conditional test to the study of lexical relationships in naturally occurring text and compare the results to those obtained using other significance tests.

Lexical Relationships

We apply significance testing to the study of two important types of lexical relationships: positive association and difference. While significance testing can be applied more generally, we limit ourselves to the study of bigrams which we define as any two consecutive words that occur together in a text.

Two consecutive words that form a bigram exhibit positive association when they occur together more often than would be expected by chance. We refer to the study of positive association as the *test for association*. We use it to determine if bigrams such as **major league** or **fine wine** exhibit positive association.

Two words are different to the extent that they are used in different contexts. The difference between two words can be studied by the *bigram difference test*, which determines how likely it is that these two words precede (or follow) the same word. Are **strong** and **powerful** equally likely to precede **tea**?

A more complete description of context provides a more complete assessment of the difference in the use of two words. The *extended bigram difference test* studies the behavior of two words in relation to all words that they immediately precede (or follow). Are **rise** and **fall** immediately preceded by approximately the same set of words?

The challenge in studying these lexical relationships is that most bigrams are relatively rare in a text regard-

less of the size of that text. This follows from the distributional tendencies of individual words and bigrams as described by Zipf's Law (Zipf 1935). Zipf found that if the frequencies of the words in a large text are ordered from most to least frequent, $(f_1, f_2, \ldots, f_m)$, these frequencies roughly obey: $f_i \propto \frac{1}{i}$. As an example, in a 132,755 word subset of the ACL/DCI Wall Street Journal corpus (Marcus, Santorini, & Marcinkiewicz 1993) there are 73,779 distinct bigrams. Of these, 81 percent occur once and 97 percent of them occur five times or less. As a result of these distributional tendencies, data samples characterizing bigrams are often skewed (i.e., comprised of both large and small counts) or sparse (i.e., contain a large number of zero counts). This kind of data violates the asymptotic approximations implicit in many traditional statistical procedures and affects our ability to study lexical relations using standard statistical techniques.

Representation of the Data

Here, we study lexical relations by observing bigrams. To represent the data in terms of a statistical model the features of each bigram are mapped to *discrete random variables*. The features of a bigram pertain either to the individual words forming the bigram (in the association test and the expanded difference test) or the presence or absence of a particular word pair that form a bigram (in the bigram difference test).

If each bigram in the data sample is characterized by two features represented by binary variables X and Y (as in the association test), then each bigram will have one of four possible classifications corresponding to the possible combinations of these variable values. The data is said to be *cross-classified* with respect to the variables X and Y. If there are I possible values for the first variable and J possible values for the second variable (as in the extended bigram difference test), then the frequency of each classification can be recorded in a rectangular table having I rows and J columns. Such a table is called an $I \times J$ *contingency table*. Contingency tables can extend to beyond two dimensions when an object is cross-classified with respect to more than two variables. However, in this paper only two dimensional tables are considered.

The *joint frequency distribution* of X and Y in a data sample is described by the counts $\{n_{ij}\}$ in the contingency table. The *marginal frequency distributions* of X and Y are the row and column totals obtained by summing the joint frequencies. The row totals are denoted by n_{i+}, the column totals by n_{+j} and the total sample size by n_{++}.

Significance Testing

A significance test seeks to assess the probability that a data sample has been taken from a population that can be described by a certain probabilistic model. The *form* of such a model identifies the relationship that exists, while the *parameters* express the uncertainty inherent in that relationship. Below we enumerate the steps in performing a significance test and discuss them in the sections that follow.

1. Select the appropriate sampling model,
2. hypothesize a population model for the data sample,
3. select a goodness of fit statistic to use in testing the fit of the model to the data sample, and
4. asses the statistical significance of the model: determine the probability that the data sample came from a population described by the model.

Sampling Models

Ideally the sampling model is established as part of the experimental design to be consistent with the analysis of the data. For our three experiments there are two applicable sampling models, each of which is discussed below. Each model assumes that the population under study has been randomly sampled through a series of Bernoulli trials.

multinomial sampling

This is the sampling model used for the association test. The overall sample size n_{++} is determined in advance. Figure 1 shows the data used to test the bigram **oil industry** for association. The variable X represents the presence or absence of **oil** in the first position of each bigram, and Y represents the presence or absence of **industry** in the second position of each bigram. We fix the total sample size n_{++} at 1,382,828 prior to the experiment.

		Y		
		industry	¬industry	totals
X	oil	n_{11}= 17	n_{12}= 229	n_{1+}= 246
	¬oil	n_{21}= 935	n_{22}= 1381647	n_{2+}= 1382582
	totals	n_{+1}=952	n_{+2}=1381876	n_{++}=1382828

Figure 1: Test for Association

product multinomial sampling

This is the sampling model used for the bigram difference test and the extended bigram difference test. It fixes not only the sample size n_{++} but also the row totals n_{i+}. This model is appropriate when two populations are sampled and the size of the sample taken from each population is fixed in advance.

The data for the bigram difference test comparing the bigrams **oil industry** and **chemical industry** is shown in Figure 2. The row totals are both fixed prior to sampling at 1,382,828 which fixes the sample size at 2,765,656. In effect, two separate multinomial samplings are performed. The first involves selecting and examining 1,382,828 bigrams to see if they are **oil industry** or not. The second inspects 1,382,828 bigrams to see if they are **chemical industry** or not.

		Y		
		yes	no	totals
X	oil industry	17	1382811	1382828
	chemical industry	10	1382818	1382828
	totals	27	2765629	2765656

Figure 2: Bigram Difference Test

Product multinomial sampling is also used in the extended bigram difference test. Figure 3 presents a portion of the data used to study the difference between the words **came** and **went** based on the context of the immediately preceding word.

		Y							
		he	it	i	···	ibm	in	but	totals
X	came	1	1	5	···	1	2	1	218
	went	0	1	0	···	0	1	0	132
	totals	1	2	5	···	1	3	1	350

Figure 3: Extended Bigram Difference Test

Here, X denotes the second word in the bigram, **came** or **went**, and Y denotes the word that precedes either **came** or **went**. Again, two multinomial samplings are performed. The first involves selecting and examining 218 bigrams having **came** as the second word. The second involves the same for 132 bigrams having **went** as the second word.

Hypothesizing a Model

The population model used to study association between two words, where the two words are represented by the binary variables X and Y, is the model for independence between X and Y (below, x and y denote the values of X and Y, respectively):

$$P(x, y) = P(x)\,P(y) \tag{1}$$

If the model for independence fits the data well as measured by its statistical significance, then one can infer from this data sample that these two words are independent in the larger population. The worse the fit, the more associated the words are judged to be. Maximum likelihood estimates of the parameters of the model for independence between two words are:

$$\hat{P}(x_i) = \frac{n_{i+}}{n_{++}} \qquad \hat{P}(y_j) = \frac{n_{+j}}{n_{++}} \tag{2}$$

Substituting these estimates into equation 1, we obtain the the following estimates for the joint probability of X and Y in the population:

$$\hat{P}(x_i, y_j) = \frac{n_{i+}}{n_{++}} \times \frac{n_{+j}}{n_{++}} \tag{3}$$

The difference between two words as characterized by the contexts that they occur in can be studied using a model specifying homogeneity under product multinomial sampling. This model is appropriate if the data sample has been generated via two multinomial samplings. The model for homogeneity specifies that these two samples come from populations that can be described by the same model. When this is true, the populations are said to be homogeneous. The model for homogeneity can also be cast as a model for independence (i.e., equation 1) by stating that the context is the same for both words and therefore independent of the word being studied.

Goodness of Fit Statistics

A goodness of fit statistic is used to measure how closely the events observed in a data sample correspond to those that would be expected if the null hypothesis were true. Evaluating this correspondence is referred to as measuring the fit of the model. In this section, we discuss three metrics that have been used to measure the fit of the models for association and difference: the likelihood ratio statistic G^2, Pearson's X^2 statistic, and the t-statistic.

power divergence family

This family of statistics was introduced in (Cressie & Read 1984) and includes the well known goodness of fit statistics X^2 and G^2. These statistics measure the divergence of observed (n_{ij}) and expected (m_{ij}) sample counts, where m_{ij} is calculated assuming that the null hypothesis is correct. From equation 3 the maximum likelihood estimates for the expected counts are:

$$m_{ij} = \hat{P}(x_i, y_j) \times n_{++} = \frac{n_{i+}\, n_{+j}}{n_{++}} \tag{4}$$

Using the above, G^2 and X^2 are calculated as:

$$G^2 = 2 \sum_{i,j} n_{ij} \log \frac{n_{ij}}{m_{ij}} \qquad X^2 = \sum_{i,j} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{5}$$

When the hypothetical population model corresponds to the actual population model, the distributions of both G^2 and X^2 converge to χ^2 as the sample size grows large. The χ^2 distribution is an *asymptotic approximation* for the distributions of G^2 and X^2. More precisely, X^2 and G^2 are approximately χ^2 distributed when the following conditions regarding the data sample hold (Read & Cressie 1988):

1. the sample size is large,
2. the number of cells in the contingency table is fixed and small relative to the sample size, and
3. the expected count under the hypothetical population model for each cell is large.

A sample is considered large when it is several times larger than the number of cells in the table. While this requirement is not a concern for 2 × 2 tables, it is for the larger two dimensional tables used in the extended bigram difference test.

When the data is sparse or skewed, the assumption regarding the expected count for each cell makes the asymptotic approximation unreliable. It is well known that G^2 and X^2 hold to a χ^2 distribution when the minimum of the expected cell counts under the null hypothesis is five. An early summary of research leading to that conclusion is found in (Cochran 1952). However, when some of the expected counts are less than five, the validity of the asymptotic approximation remains an active point of research. A summary of recent work can be found in (Read & Cressie 1988). There it is shown that when the individual cell counts in a table are all approximately equal, the asymptotic approximation holds for expected counts as small as one. Unfortunately, the counts in 2×2 tables representing bigram data are almost always heavily skewed. In the test for association, n_{11} is usually a very small count while n_{22} approaches the size of the sample. In the bigram difference test, the counts n_{11} and n_{21} are usually small while n_{12} and n_{22} approach the marginal totals.

Failure to meet any of the three conditions described above can mean that the actual distribution of these goodness of fit statistics is far from the χ^2 approximation. In that case, significance values assigned based on the χ^2 approximation can be incorrect. We present an approach to assigning statistical significance to a goodness of fit statistic that does not use the asymptotic approximation to the distribution of that statistic. Before describing this we discuss an approach to evaluating the significance of lexical relations using the normal approximation to the t-statistic.

the t-statistic

The t-statistic is used in the *t-test*, a test to determine how likely it is that either a single population mean (a one-sample t-test) or the difference between two population means (a two-sample t-test) equates to a hypothesized value. In both tests, the t-statistic is the difference between the observed value and the hypothesized value scaled by the estimated variance. These two formulations of the t-test are utilized in (Church *et al.* 1991) to perform the test for association and the bigram difference test. In both formulations a first-order approximation is used to equate the variance of a series of Bernoulli trials recording the presence or absence of a single bigram with the observed relative frequency of the bigram in those trials (i.e., the sample mean).

According to (Church *et al.* 1991), lexical association can be evaluated using a one-sample t-test where the t-statistic compares the observed relative frequency of a bigram to the expected relative frequency under the model for independence. The t-statistic measuring the association of a bigram $x_i\, y_j$ is formulated as:

$$t \approx \frac{\hat{P}(x_i, y_j) - \hat{P}(x_i)\hat{P}(y_j)}{\sqrt{\frac{\hat{P}(x_i,y_j)}{n_{++}}}} = \frac{n_{11} - m_{11}}{\sqrt{n_{11}}} \qquad (6)$$

Significance is assigned to the t-statistic using the t-distribution, which converges to the standard normal distribution as the sample size grows. This approach to assigning significance is based on the assumption that the sample means are normally distributed. When this assumption does not hold, the significance assigned can be incorrect.

As shown in (Church *et al.* 1991), the bigram difference test can be evaluated using a two-sample t-test where the t-statistic compares the relative frequencies of two bigrams. However, it can also be shown that a two-sample t-test is identical to Pearson's X^2 test[1] when applied to a 2×2 contingency table (Fisher 1968). Thus, the two-sample t-test applied to 2×2 contingency tables is appropriate only when the three conditions described in the previous section are met.

Assessing Statistical Significance

If the statistic used to evaluate model fit has a known distribution when the model is correct, that distribution can be used to assign statistical significance. The statistical significance of a model is the likelihood that the data sample was randomly selected from a population described by the model; it is equal to the likelihood that the value calculated for that statistic came from the distribution of statistic values that occurs when the null hypothesis is true. As discussed so far, this distribution can be approximated when certain assumptions hold. The problem is that these assumptions are frequently violated by the data found in NLP.

An alternative to using an asymptotic approximation to the distribution of a goodness of fit test statistic is to define its exact distribution. There are two ways to define the exact distribution of a test statistic: (1) enumerate all elements of that distribution as in Fisher's Exact Test (Fisher 1966), or (2) sample from that distribution using a Monte Carlo sampling scheme (Ripley 1987). The freely available software package CoCo (Badsberg 1995) implements the Monte Carlo sampling scheme described in (Kreiner 1987) and summarized below:

1. Generate a random sample of comparable tables from the model being tested. A comparable table is one having the same marginal totals as the observed table. (Patefield 1981) presents an algorithm for $I \times J$ tables.
2. Calculate the value of the test statistic for each of the tables in the random sample.
3. The significance value of the observed table is approximated by the number of random tables whose test statistic values are greater than the test statistic value of the observed table divided by the total number of random tables.

[1] Pearson's X^2 test is a significance test in which the X^2 statistic is used to measure model fit and the χ^2 distribution is used to assign significance as described in the previous section.

The advantage of using the exact distribution of a goodness of fit test statistic (as opposed to the asymptotic approximation of the distribution of that statistic) when assessing statistical significance is that the assessment is free of assumptions. Therefore, it is accurate for skewed and sparse data samples.

Experiments and Results

In these experiments we study the lexical relationships described previously by comparing the significance values computed using the normal approximation to the distribution of the t-statistic, the χ^2 approximation to the distribution of both G^2 and X^2, and the exact conditional distribution of G^2 as defined by CoCo.

Our objective is to demonstrate that these tests assign different significance values for the same data. When this occurs, the assumptions required by certain of these tests are being violated. In that case, the significance values assigned using the exact conditional distribution of G^2 are more reliable because there are no restrictions on the nature of the data required for this test.

association test

The test for association was performed using a 1.38 million word subset of the ACL/DCI Wall Street Journal corpus. We study the associations formed by the word **industry** by observing bigrams of the form **<word> industry**, where **<word>** denotes a word spelling form.

Figure 5 displays a subset of the bigrams found in the corpus. Shown are the significance values assigned by the various tests for association between the word in column one and **industry**. A significance value of .000 implies that this data shows no evidence of independence; the likelihood of having randomly selected this data sample from a population where these words are independent is zero. A significance value of 1.00 indicates that the data sample is exactly what would be expected from a population where the words are independent and there is no reason to doubt that this sample was drawn from such a population.

In general, the data in Figure 5 confirms the findings of (Dunning 1993); based on a comparison to the exact conditional test, the χ^2 approximation to the distribution of G^2 is found to be more reliable than the normal approximation to the distribution of the t-statistic and the χ^2 approximation to the distribution of Pearson's X^2. (Dunning 1993) showed that when the minimum of the expected values in a 2×2 table is one, then G^2 holds more closely to χ^2 than X^2. However, it is frequently the case that the minimum expected value is much less than one. In this case, the validity of the asymptotic approximation of G^2 is an open question (Read & Cressie 1988). For example, in the case of the bigram **new industry** the significance values assigned by each test are different. Such cases are indicative of a failure to meet the requirements of the asymptotic approximation and should be analyzed using the exact conditional test.

<word>	exact	$G^2 \sim \chi^2$	$X^2 \sim \chi^2$	$t \sim N$
or	1.00	.761	.768	.994
million	1.00	.795	.800	.995
new	.382	.191	.259	.976
white	.210	.243	.116	.967
power	.209	.242	.114	.967
energy	.146	.156	.034	.956
community	.112	.114	.011	.947
this	.108	.163	.113	.967
and	.098	.097	.078	.964
recent	.086	.105	.030	.955
aircraft	.084	.080	.002	.936
farm	.060	.054	.000	.921
surprised	.037	.031	.000	.897
broadcast	.029	.024	.000	.882
carpet	.012	.008	.000	.811
glass	.001	.000	.000	.783
dominant	.000	.000	.000	.742
s&l	.000	.000	.000	.692
in	.000	.000	.000	.925
loan	.000	.000	.000	.769
basic	.000	.000	.000	.626
some	.000	.000	.000	.804
petroleum	.000	.000	.000	.557

Figure 5: Test for association **<word> industry**

difference in context

A 928,000 word subset of the ACL/DCI Wall Street Journal corpus was used in both the bigram difference test and the extended bigram difference test.

bigram difference test In this experiment bigrams of the form **<word> came** and **<word> went** are compared. A subset of the bigrams found in the corpus are shown in Figure . Here, a significance of 1.00 corresponds to a 100% probability that **<word> came** and **<word> went** behave in a similar manner in the larger population as judged from the counts in the data sample. In addition to the significance values, the frequency of **<word> came** is shown in column n_{11} and the frequency of **<word> went** is shown in column n_{21}. In this experiment the most noticeable differences between the significance values assigned by the various tests occur when $|n_{11} - n_{21}|$ is equal to one. The exact conditional test assigns a significance of 1.00 while the other tests assign lower values. The assignment of 1.00 is reasonable since the distinction between **<word> came** and **<word> went** is as small as possible. When **<word>** is observed an odd number of times it can not occur the same number of times with **came** as it does with **went**.

Note that the significance assigned by the asymptotic approximations to the t-statistic and X^2 are iden-

word	n_{11}	n_{21}	exact	$G^2 \sim \chi^2$	$t \sim N$ $X^2 \sim \chi^2$
he	6	6	1.00	1.00	1.00
we	4	4	1.00	1.00	1.00
bonds	1	1	1.00	1.00	1.00
it	7	6	1.00	.781	.782
she	3	2	1.00	.654	.655
i	2	3	1.00	.654	.655
corp	1	2	1.00	.560	.564
agents	0	1	1.00	.239	.317
accident	1	0	1.00	.239	.317
and	1	3	.625	.306	.317
revenue	3	1	.625	.306	.317
approval	2	0	.500	.096	.157
ibm	0	2	.500	.096	.157
who	7	3	.344	.200	.206
notice	3	0	.250	.041	.083
they	5	1	.219	.088	.102
what	0	4	.125	.018	.045
then	6	1	.125	.046	.059
action	5	0	.062	.008	.025
which	7	0	.016	.002	.008

Figure 6: Bigram difference test
<word> came vs. **<word> went**

tical. This demonstrates the equivalence of Pearson's X^2 test and the bigram difference t-test.

extended bigram difference test This experiment compares **<word> came** and **<word> went** where **<word>** has as possible values any of the spelling forms that precede **came** or **went**. This data is represented in a 2×237 contingency table, a portion of which is shown in Figure 3. This table clearly violates the assumption that the number of cells in the table be less than the sample size. In this case the sample size is 350 and the number of cells in the table is 574.

The significance values assigned using the χ^2 approximation to the distributions of X^2 and G^2 are .292 and .000, respectively. This difference is indicative of violations of the assumptions supporting the use of the χ^2 approximation. The t-test was not included in this experiment because its application to tables larger than 2×2 is at best problematic.

The exact conditional test, as performed by CoCo, assigns a significance of .0670 to the extended bigram difference test for **came** and **went**. This value is reliable since it is arrived at without making any assumptions about the nature of the data being tested.

Conclusions

A number of standard significance tests have been applied to the study of lexical relationships. All of these tests use asymptotic approximations to the distribution of the test statistic, and therefore are not appropriate for the sparse and skewed data typically found in NLP. In this paper, we have described a test that can be used to accuratcly assess the significance of a population model from a sample of NLP data. This test, an exact conditional test, assigns significance by generating the exact distribution of the test statistic using a Monte Carlo sampling scheme. This test can be performed using a freely available software package called CoCo (Badsberg 1995).

Acknowledgments

This research was supported by the Office of Naval Research under grant number N00014-95-1-0776.

References

Badsberg, J. 1995. *An Environment for Graphical Models.* Ph.D. Dissertation, Aalborg University.

Church, K.; Gale, W.; Hanks, P.; and Hindle, D. 1991. Using statistics in lexical analysis. In Zernik, U., ed., *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cochran, W. 1952. The χ^2 test of goodness of fit. *Annals of Mathematical Statistics* 23:315–345.

Cressie, N., and Read, T. 1984. Multinomial goodness-of-fit tests. *Journal of the Royal Statistics Society Series B* 46:440–464.

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

Fisher, R. 1966. *The Design of Experiments.* New York, NY: Hafner, eighth edition.

Fisher, R. 1968. *Statistical Methods for Research Workers.* New York, NY: Hafner, thirteenth edition.

Kreiner, S. 1987. Analysis of multidimensional contingency tables by exact conditional tests: Techniques and strategies. *Scandinavian Journal of Statistics* 14:97–112.

Marcus, M.; Santorini, B.; and Marcinkiewicz, M. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2):313–330.

Patefield, W. 1981. An efficient method of generating random R × C tables with given row and column totals. *Applied Statistics* 30:91–97.

Read, T., and Cressie, N. 1988. *Goodness-of-fit Statistics for Discrete Multivariate Data.* New York, NY: Springer-Verlag.

Ripley, B. 1987. *Stochastic Simulation.* New York, NY: John Wiley.

Zipf, G. 1935. *The Psycho-Biology of Language.* Boston, MA: Houghton Mifflin.