More accurate tests for the statistical significance of result differences [*]

Alexander Yeh
Mitre Corp.
202 Burlington Rd.
Bedford, MA 01730
USA
asy@mitre.org

Abstract

Statistical significance testing of differences in values of metrics like recall, precision and balanced F-score is a necessary part of empirical natural language processing. Unfortunately, we find in a set of experiments that many commonly used tests often underestimate the significance and so are less likely to detect differences that exist between different techniques. This underestimation comes from an independence assumption that is often violated. We point out some useful tests that do not make this assumption, including computationally-intensive randomization tests.

1 Introduction

In empirical natural language processing, one is often testing whether some new technique produces improved results (as measured by one or more metrics) on some test data set when compared to some current (baseline) technique. When the results are better with the new technique, a question arises as to whether these result differences are due to the new technique actually being better or just due to chance. Unfortunately, one usually cannot directly answer the question "what is the probability that the new technique is better given the results on the test data set":

P(new technique is better | test set results)

But with statistics, one can answer the following proxy question: if the new technique was actually no different than the old technique (the null hypothesis), what is the probability that the results on the test set would be at least this skewed in the new technique's favor (Box et al., 1978, Sec. 2.3)? That is, what is

P(test set results at least this skewed
in the new technique's favor
| new technique is no different than the old)

If the probability is small enough (5% often is used as the threshold), then one will reject the null hypothesis and say that the differences in the results are "statistically significant" at that threshold level.

This paper examines some of the possible methods for trying to detect statistically significant differences in three commonly used metrics: recall, precision and balanced F-score. Many of these methods are found to be problematic in a set of experiments that are performed. These methods have a tendency to underestimate the significance of the results, which tends to make one believe that some new technique is no better than the current technique even when it is.

This underestimate comes from these methods assuming that the techniques being compared produce independent results when in our experiments, the techniques being compared tend to produce positively correlated results.

To handle this problem, we point out some statistical tests, like the matched-pair t, sign and Wilcoxon tests (Harnett, 1982, Sec. 8.7 and 15.5), which do not make this assumption. One can use these tests on the recall metric, but the precision and balanced F-score metric have too complex a form for these tests. For such complex metrics, we use a compute-intensive randomization test (Cohen, 1995, Sec. 5.3), which also avoids this independence assumption.

[*] This paper reports on work performed at the MITRE Corporation under the support of the MITRE Sponsored Research Program. Warren Greiff, Lynette Hirschman, Christine Doran, John Henderson, Kenneth Church, Ted Dunning, Wessel Kraaij, Mitch Marcus and an anonymous reviewer provided helpful suggestions. Copyright ©2000 The MITRE Corporation. All rights reserved.

The next section describes many of the standard tests used and their problem of assuming certain forms of independence. The first subsection describes tests where this assumption appears in estimating the standard deviation of the difference between the techniques' results. The second subsection describes using contingency tables and the χ^2 test. Following this is a section on methods that do not make this independence assumption. Subsections in turn describe some analytical tests, how they can apply to recall but not precision or the F-score, and how to use randomization tests to test precision and F-score. We conclude with a discussion of dependencies within a test set's instances, a topic that we have yet to deal with.

2 Tests that assume independence between compared results

2.1 Finding and using the variance of a result difference

For each metric, after determining how well a new and current technique performs on some test set according to that metric, one takes the difference between those results and asks "is that difference significant?"

A way to test this is to expect no difference in the results (the null hypothesis) and to ask, assuming this expectation, how unusual are these results? One way to answer this question is to assume that the difference has a normal or t distribution (Box et al., 1978, Sec. 2.4). Then one calculates the following:

$$(d - E[d])/s_d = d/s_d \tag{1}$$

where $d = x_1 - x_2$ is the difference found between x_1 and x_2, the results for the new and current techniques, respectively. $E[d]$ is the expected difference (which is 0 under the null hypothesis) and s_d is an estimate of the standard deviation of d. Standard deviation is the square root of the variance, a measure of how much a random variable is expected to vary. The results of equation 1 are compared to tables (c.f. in Box et al. (1978, Appendix)) to find out what the chances are of equaling or exceeding the equation 1 results if the null hypothesis were true. The larger the equation 1 results, the more unusual it would be under the null hypothesis.

A complication of using equation 1 is that one usually does not have s_d, but only s_1 and s_2, where s_1 is the estimate for x_1's standard deviation and similarly for s_2. How does one get the former from the latter? It turns out that (Box et al., 1978, Ch. 3)

$$\sigma_d^2 = \sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2$$

where σ_i is the true standard deviation (instead of the estimate s_i) and ρ_{12} is the correlation coefficient between x_1 and x_2. Analogously, it turns out that

$$s_d^2 = s_1^2 + s_2^2 - 2r_{12}s_1s_2 \tag{2}$$

where r_{12} is an estimate for ρ_{12}. So not only does σ_d (and s_d) depend on the properties of x_1 and x_2 in isolation, it also depends on how x_1 and x_2 interact, as measured by ρ_{12} (and r_{12}). When x_1 and x_2 are independent, $\rho_{12} = 0$, and then $\sigma_d = \sqrt{\sigma_1^2 + \sigma_2^2}$ and analogously, $s_d = \sqrt{s_1^2 + s_2^2}$. When ρ_{12} is positive, x_1 and x_2 are positively correlated: a rise in x_1 or x_2 tends to be accompanied by a rise in the other result. When ρ_{12} is negative, x_1 and x_2 are negatively correlated: a rise in x_1 or x_2 tends to be accompanied by a decline in the other result. $-1 \leq \rho_{12} \leq 1$ (Larsen and Marx, 1986, Sec. 10.2).

The assumption of independence is often used in formulas to determine the statistical significance of the difference $d = x_1 - x_2$. But how accurate is this assumption? One might expect some positive correlation from both results coming from the same test set. One may also expect some positive correlation when either both techniques are just variations of each other[1] or both techniques are trained on the same set of training data (and so are missing the same examples relative to the test set).

This assumption was tested during some experiments for finding grammatical relations (subject, object, various types of modifiers, etc.). The metric used was the fraction of the relations of interest in the test set that were *recall*ed (found) by some technique. The relations of interest were various subsets of the 748 relation instances in that test set. An example subset is all the modifier relations. Another subset is just that of all the time modifier relations.

[1] These variations are often designed to usually behave in the same way and only differ in just a few cases.

First, two different techniques, one memory-based and the other transformation-rule based, were trained on the same training set, and then both tested on that test set. Recall comparisons were made for ten subsets of the relations and the r_{12} was found for each comparison. From Box et al. (1978, Ch. 3)

$$r_{12} = \sum_k (y_{1k} - \overline{y}_1)(y_{2k} - \overline{y}_2)/(s_1 s_2 (n-1))$$

where $y_{ik} = 1$ if the ith technique recalls the kth relation and $= 0$ if not. n is the number of relations in the subset. $\overline{y}_i$ and s_i are mean and standard deviation estimates (based on the y_{ik}'s), respectively, for the ith technique.

For the ten subsets, only one comparison had a r_{12} close to 0 (It was -0.05). The other nine comparisons had r_{12}'s between 0.29 and 0.53. The ten comparison median value was 0.38.

Next, the transformation-rule based technique was run with different sets of starting conditions and/or different, but overlapping, subsets of the training set. Recall comparisons were made on the same test data set between the different variations. Many of the comparisons were of how well two variations recalled a particular subset of the relations. A total of 40 comparisons were made. The r_{12}'s on all 40 were positive. 3 of the r_{12}'s were in the 0.20-0.30 range. 24 of the r_{12}'s were in the 0.50-0.79 range. 13 of the r_{12}'s were in the 0.80-1.00 range.

So in our experiments, we were usually comparing positively correlated results. How much error is introduced by assuming independence? An easy-to-analyze case is when the standard deviations for the results being compared are the same.[2] Then equation 2 reduces to $s_d = s\sqrt{2(1-r_{12})}$, where $s = s_1 = s_2$. If one assumes the results are independent (assume $r_{12} = 0$), then $s_d = s\sqrt{2}$. Call this value s_{d-ind}. As r_{12} increases in value, s_d decreases:

r_{12}	s_d	$(s_{d-ind})/s_d$
0.38	$0.787(s_{d-ind})$	1.27
0.50	$0.707(s_{d-ind})$	1.41
0.80	$0.447(s_{d-ind})$	2.24

The rightmost column above indicates the magnitude by which erroneously assuming independence (using s_{d-ind} in place of s_d) will increase the standard deviation estimate. In equation 1, s_d forms the denominator of the ratio d/s_d. So erroneously assuming independence will mean that the numerator d, the difference between the two results, will need to increase by that same factor in order for equation 1 to have the same value as without the independence assumption. Since the value of that equation indicates the statistical significance of d, assuming independence will mean that d will have to be larger than without the assumption to achieve the same apparent level of statistical significance. From the table above, when $r_{12} = 0.50$, d will need to be about 41% larger. Another way to look at this is that assuming independence will make the same value of d appear less statistically significant.

The common tests of statistical significance use this assumption. The test known as the t (Box et al., 1978, Sec. 4.1) or two-sample t (Harnett, 1982, Sec. 8.7) test does. This test uses equation 1 and then compares the resulting value against the t distribution tables. This test has a complicated form for s_d because:

1. x_1 and x_2 can be based on differing numbers of samples. Call these numbers n_1 and n_2 respectively.
2. In this test, the x_i's are each an n_i sample average of another variable (call it y_i). This is important because the s_i's in this test are standard deviation estimates for the y_i's, not the x_i's. The relationship between them is that s_i for y_i is the same as $(\sqrt{n_i})s_i$ for x_i.
3. The test itself assumes that y_1 and y_2 have the same standard deviation (call this common value s). The denominator estimates s using a weighted average of s_1 and s_2. The weighting is based on n_1 and n_2.

From Harnett (1982, Sec. 8.7), the denominator

$$s_d = \sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\frac{n_1+n_2}{n_1 n_2}\right)}$$

When $n_1 = n_2$ (call this common value n), s_1 and s_2 will be given equal weight, and s_d simplifies to $\sqrt{(s_1^2 + s_2^2)/n}$. Making the substitution described above of $s_i\sqrt{n_i}$ for s_i leads to an s_d of

[2] This is actually roughly true in the comparisons made, and is assumed to be true in many of the standard tests for statistical significance.

$\sqrt{s_1^2 + s_2^2}$, the form we had earlier for using the independence assumption.

Another test that both makes this assumption and uses a form of equation 1 is a test for binomial data (Harnett, 1982, Sec. 8.11) which uses the "fact" that binomial distributions tend to approximate normal distributions. In this test, the x_i's being compared are the fraction of the items of interest that are recovered by the ith technique. In this test, the denominator s_d of equation 1 also has a complicated form, both due to the reasons mentioned for the t test above and to the fact that with a binomial distribution, the standard deviation is a function of the number of samples and the mean value.

2.2 Using contingency tables and χ^2 to test precision

A test that does *not* use equation 1 but still makes an assumption of independence between x_1 and x_2 is that of using contingency tables with the chi-squared (χ^2) distribution (Box et al., 1978, Sec. 5.7). When the assumption is valid, this test is good for comparing differences in the *precision* metric. Precision is the fraction of the items "found" by some technique that are actually of interest. Precision $= R/(R+S)$, where R is the number of items that are of interest and are Recalled (found) by the technique, and S is the number of items that are found by the technique that turn out to be Spurious (not of interest). One can test whether the precision results from two techniques are different by using a 2×2 contingency table to test whether the ratio R/S is different for the two techniques. One makes the latter test by seeing if the assumption that the ratios for the two techniques are the same (the null hypothesis) leads to a statistically significant result when using a χ^2 distribution with one degree of freedom. A 2×2 table has 4 cells. The top 2 cells are filled with the R and S of one technique and the bottom 2 cells get the R and S of the other technique. In this test, the value in each cell is assumed to have a Poisson distribution. When the cell values are not too small, these Poisson distributions are approximately Normal (Gaussian). As a result, when the cell values are independent, summing the normalized squares of the difference between each cell and its expected value leads to a χ^2 distribution (Box et al., 1978, Sec. 2.5-2.6).

How well does this test work in our experiments? Precision is a non-linear function of two random variables R and S, so we did not try to estimate the correlation coefficient for precision. However, we can easily estimate the correlation coefficients for the R's. They are the r_{12}'s found in section 2.1. As that section mentions, the r_{12}'s found are just about always positive. So at least in our experiments, the R's are not independent, but are positively correlated, which violates the assumptions of the test.

An example of how this test behaves is the following comparison of the precision of two different methods at finding the modifier relations using the same training and test set. The correlation coefficient estimate for R is 0.35 and the data is

Method	R	S	Precision
1	47	48	49%
2	25	14	64%

Placing the R and S values into a 2×2 table leads to a χ^2 value of 2.38.[3] At 1 degree of freedom, the χ^2 tables indicate that if the null hypothesis were true, there would be a 10% to 20% chance of producing a χ^2 value at least this large. So according to this test, this much of an observed difference in precision would not be unusual if no actual difference in the precision exists between the two methods.

This test assumes independence between the R values. When we use a 2^{20} (=1048576) trial approximate randomization test (section 3.3), which makes no such assumptions, then we find that this latter test indicates that under the null hypothesis, there is less than a 4% chance of producing a difference in precision results as large as the one observed. So this latter test indicates that this much of an observed difference in precision would be unusual if no actual difference in the precision exists between the two methods.

It should be mentioned that the manner of testing here is slightly different than the manner in the rest of this paper. The χ^2 test looks at the square of the difference of two results, and rejects the null hypothesis (the compared techniques are the same) when this square is

[3] We do *not* use Yate's adjustment to compensate for the numbers in the table being integers. Doing so would have made the results even worse.

large, whether the largeness is caused by the new technique producing a much better result than the current technique or vice-versa. So to be fair, we compared the χ^2 results with a two-sided version of the randomization test: estimate the likelihood that the observed magnitude of the result difference would be matched or exceeded (regardless of which technique produced the better result) under the null hypothesis. A one-sided version of the test, which is comparable to what we use in the rest of the paper, estimates the likelihood of a different outcome under the null hypothesis: that of matching or exceeding the difference of how much better the new (possibly better) technique's observed result is than the current technique's observed result. In the above scenario, a one-sided test produces a 3% figure instead of a 4% figure.

3 Tests without that independence assumption

3.1 Tests for matched pairs

At this point, one may wonder if all statistical tests make such an independence assumption. The answer is no, but those tests that do not measure how much two techniques interact do need to make some assumption about that interaction and typically, that assumption is independence. Those tests that notice in some way how much two techniques interact can use those observations instead of relying on assumptions.

One way to measure how two techniques interact is to compare how similarly the two techniques react to various parts of the test set. This is done in the matched-pair t test (Harnett, 1982, Sec. 8.7). This test finds the difference between how techniques 1 and 2 perform on each test set sample. The t distribution and a form of equation 1 are used. The null hypothesis is still that the numerator d has a 0 mean, but d is now the sum of these difference values (divided by the number of samples), instead of being $x_1 - x_2$. Similarly, the denominator s_d is now estimating the standard deviation of these difference values, instead of being a function of s_1 and s_2. This means for example, that even if the values from techniques 1 and 2 vary on different test samples, s_d will now be 0 if on each test sample, technique 1 produces a value that is the same constant amount more than the value from technique 2.

Two other tests for comparing how two techniques perform by comparing how well they perform on each test sample are the sign and Wilcoxon tests (Harnett, 1982, Sec. 15.5). Unlike the matched-pair t test, neither of these two tests assume that the sum of the differences has a normal (Gaussian) distribution. The two tests are so-called nonparametric tests, which do not make assumptions about how the results are distributed (Harnett, 1982, Ch. 15).

The sign test is the simplier of the two. It uses a binomial distribution to examine the number of test samples where technique 1 performs better than technique 2 versus the number where the opposite occurs. The null hypothesis is that the two techniques perform equally well.

Unlike the sign test, the Wilcoxon test also uses information on how large a difference exists between the two techniques' results on each of the test samples.

3.2 Using the tests for matched-pairs

All three of the matched-pair t, sign and Wilcoxon tests can be applied to the recall metric, which is the fraction of the items of interest in the test set that a technique recalls (finds). Each item of interest in the test data serves as a test sample. We use the sign test because it makes fewer assumptions than the matched-pair t test and is simplier than the Wilcoxon test. In addition, the fact that the sign test ignores the size of the result difference on each test sample does not matter here. With the recall metric, each sample of interest is either found or not by a technique. There are no intermediate values.

While the three tests described in section 3.1 can be used on the recall metric, they cannot be straightforwardly used on either the precision or balanced F-score metrics. This is because both precision and F-score are more complicated nonlinear functions of random variables than recall. In fact both can be thought of as non-linear functions involving recall. As described in Section 2.2, precision $= R/(R+S)$, where R is the number of items that are of interest that are *recall*ed by a technique and S is the number of items (found by a technique) that are not of interest. The balanced F-score $= 2ab/(a+b)$, where a is recall and b is precision.

3.3 Using randomization for precision and F-score

A class of technique that can handle all kinds of functions of random variables without the above problems is the computationally-intensive randomization tests (Noreen, 1989, Ch. 2) (Cohen, 1995, Sec. 5.3). These tests have previously used on such functions during the "message understanding" (MUC) evaluations (Chinchor et al., 1993). The randomization test we use is like a randomization version of the paired sample (matched-pair) t test (Cohen, 1995, Sec. 5.3.2). This is a type of stratified shuffling (Noreen, 1989, Sec. 2.7). When comparing two techniques, we gather-up all the responses (whether actually of interest or not) produced by one of the two techniques when examining the test data, but not both techniques. Under the null hypothesis, the two techniques are not really different, so any response produced by one of the techniques could have just as likely come from the other. So we shuffle these responses, reassign each response to one of the two techniques (equally likely to either technique) and see how likely such a shuffle produces a difference (new technique minus old technique) in the metric(s) of interest (in our case, precision and F-score) that is at least as large as the difference observed when using the two techniques on the test data.

n responses to shuffle and assign[4] leads to 2^n different ways to shuffle and assign those responses. So when n is small, one can try each of the different shuffles once and produce an exact randomization. When n gets large, the number of different shuffles gets too large to be exhaustively evaluated. Then one performs an approximate randomization where each shuffle is performed with random assignments.

For us, when $n \leq 20$ ($2^n \leq 1048576$), we use an exact randomization. For $n > 20$, we use an approximate randomization with 1048576 shuffles. Because an approximate randomization uses random numbers, which both lead to occasional unusual results and may involve using a not-so-good pseudo-random number generator[5], we perform the following checks:

- We run the 1048576 shuffles a second time and compare the two sets of results.
- We also use the same shuffles to calculate the statistical significance for the recall metric, and compare this significance value with the significance value found for recall analytically by the sign test.

An example of using randomization is to compare two different methods on finding modifier relations in the same test set. The results on the test set are:

Method	Recall	Precision	F-score
I	45.6%	49.5%	47.5%
II	24.3%	64.1%	35.2%

Two questions being tested are whether the apparent improvement in recall and F-score from using method I is significant. Also being tested is whether the apparent improvement in precision from using method II is significant.

In this example, there are 103 relations that should be found (are of interest). Of these, 19 are recalled by both methods, 28 are recalled by method I but not II, and 6 are recalled by II but not I. The correlation coefficient estimate between the methods' recalls is 0.35. In addition, 5 spurious (not of interest) relations are found by both methods, with method I finding an additional 43 spurious relationships (not found by method II) and method II finding an additional 9 relationships.

There are a total of 28+6+43+9=86 relations that are found (whether of interest or not) by one method, but not the other. This is too many to perform an exact randomization, so a 1048576 trial approximate randomization is performed.

In 96 of these trials, method I's recall is greater than method II's recall by at least (45.6%−24.3%). Similarly, in 14794 of the trials, the F-score difference is at least (47.5%−35.2%). In 25770 of the trials, method II's precision is greater than method I's precision by at least (64.1%−49.5%). From (Noreen, 1989, Sec. 3A.3), the significance level (probability under the null hypothesis) is at most $(nc + 1)/(nt + 1)$, where nc is the number of trials that meet the criterion and nt is the number of trials. So for recall, the significance level is at most (96+1)/(1048576+1) =0.00009.

[4] Note that responses produced by both or neither techniques do not need to be shuffled and assigned.

[5] One example is the RANDU routine on the IBM360 (Forsythe et al., 1977, Sec. 10.1).

Similarly, for F-score, the significance level is at most 0.014 and for precision, the level is at most 0.025. A second 1048576 trial produces similar results, as does a sign test on recall. Thus, we see that all three differences are statistically significant.

4 The future: handling inter-sample dependencies

An assumption made by all the methods mentioned in this paper is that the members of the test set are all independent of one another. That is, knowing how a method performs on one test set sample should not give any information on how that method performs on other test set samples. This assumption is not always true.

Church and Mercer (1993) give some examples of dependence between test set instances in natural language. One type of dependence is that of a lexeme's part of speech on the parts of speech of neighboring lexemes (their section 2.1). Similar is the concept of *collocation*, where the probability of a lexeme's appearance is influenced by the lexemes appearing in nearby positions (their section 3). A type of dependence that is less local is that often, a content word's appearance in a piece of text greatly increases the chances of that same word appearing later in that piece of text (their section 2.3).

What are the effects when some dependency exists? The expected (average) value of the instance results will stay the same. However, the chances of getting an unusual result can change. As an example, take five flips of a fair coin. When no dependencies exist between the flips, the chances of the extreme result that all the flips land on a particular side is fairly small ($(1/2)^5 = 1/32$). When the flips are positively correlated, these chances increase. When the first flip lands on that side, the chances of the other four flips doing the same are now each greater than 1/2.

Since statistical significance testing involves finding the chances of getting an unusual (skewed) result under some null hypothesis, one needs to determine those dependencies in order to accurately determine those chances. Determining the effect of these dependencies is something that is yet to be done.

5 Conclusions

In empirical natural language processing, one is often comparing differences in values of metrics like recall, precision and balanced F-score. Many of the statistics tests commonly used to make such comparisons assume the independence between the results being compared. We ran a set of natural language processing experiments and found that this assumption is often violated in such a way as to understate the statistical significance of the differences between the results. We point out some analytical statistics tests like matched-pair t, sign and Wilcoxon tests, which do not make this assumption and show that they can be used on a metric like recall. For more complicated metrics like precision and balanced F-score, we use a compute-intensive randomization test, which also avoids this assumption. A next topic to address is that of possible dependencies between test set samples.

References

G. Box, W. Hunter, and J. Hunter. 1978. *Statistics for experimenters.* John Wiley and Sons.

N. Chinchor, L. Hirschman, and D. Lewis. 1993. Evaluating message understanding systems: an analysis of the third message understanding conference (muc-3). *Computational Linguistics*, 19(3).

K. Church and R. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

P. Cohen. 1995. *Empirical Methods for Artificial Intelligence.* MIT Press, MA, USA.

G. Forsythe, M. Malcolm, and C. Moler. 1977. *Computer methods for mathematical computations.* Prentice-Hall, NJ, USA.

D. Harnett. 1982. *Statistical Methods.* Addison-Wesley Publishing Co., 3rd edition.

R. Larsen and M. Marx. 1986. *An Introduction to Mathematical Statistics and Its Applications.* Prentice-Hall, NJ, USA, 2nd edition.

E. Noreen. 1989. *Computer-intensive methods for testing hypotheses: an introduction.* John Wiley and Sons, Inc.