THE STATISTICAL SIGNIFICANCE OF THE MUC-4 RESULTS

Nancy Chinchor, Ph.D.
Science Applications International Corporation
10260 Campus Point Drive, M/S A2-F
San Diego, CA 92121
chinchor@esosun.css.gov
(619) 458-2614

INTRODUCTION

The MUC-4 scores of recall, precision, and the F-measures are used to measure the performance of the participating systems. The differences in the scores between any two systems may be due to chance or may be due to a significant difference between the two systems. To rule out the possibility that the difference is due to chance, statistical hypothesis testing is used. The method of hypothesis testing used is a computationally-intensive method known as approximate randomization. The method and the statistical significance of the results for the two MUC-4 test sets, TST3 and TST4, will be discussed in this paper.

In our hypothesis testing, our objective was to determine whether a system is characteristically different from another system. This was achieved by comparing two systems to see if their actual difference in performance stands out in comparison with the results for random combinations of their scores. If their actual difference stands out, then we know that this difference could not have arisen by chance.

ELEMENTS OF HYPOTHESIS TESTING

A key element in hypothesis testing is obviously the hypothesis. Statistics can be used to reject a hypothesis. In statistical hypothesis testing, it is important to compose the hypothesis in such a way as to be able to use statistics to reject it and to thereby be able to conclude that something of interest is true. The hypothesis formulated to be rejected is called the *null hypothesis*. In Table 1, some elementary terms are defined, including the *null hypothesis*.

We are interested in determining whether two systems are significantly different in their performance on a MUC-4 test set. To conclude that two systems are significantly different, we would formulate null hypotheses of the following form to be tested for each test set:

The absolute value of the difference between System X's overall recall (precision, F-measure) score for the data extraction task and System Y's overall recall (precision, F-measure) score for the data extraction task is approximately equal to zero.

If this null hypothesis can be rejected, then we can conclude that the systems are significantly different.

Another key element in hypothesis testing found within the null hypothesis is the test statistic. A *test statistic* is a function which can be applied to a set of sample data to produce a single numerical value. A simple example of a test statistic is recall, which is a function of the number correct, the number partially correct, and the number of possible fills. The test statistic we used in our hypothesis testing is the absolute value of the difference in recall, precision, or F-measure of systems. *Observations* are instances of a set of random variables. An example of an observation is the four-tuple for a MUC-4 message consisting of the number possible, actual, correct, and partially correct. This four-tuple plays an important role in our application of the approximate randomization method.

☐ Null hypothesis	<p>The hypothesis that a relationship of interest is not present.</p> <p>Examples (informal):</p> <ol style="list-style-type: none"> 1) system X and system Y do not differ in <i>recall</i> 2) system X and system Y do not differ in <i>precision</i> 3) system X and system Y do not differ in <i>F-measure</i> for equal weighting of <i>recall</i> and <i>precision</i>
☐ Test statistic	<p>A function which can be applied to a set of sample data to produce a single numerical value.</p> <p>Examples: <i>recall, precision, F-measure, difference in recall</i></p>
☐ Observations	<p>Instances of values of a set of random variables.</p> <p>Example: <i>number possible, actual, correct, partially correct</i></p>
☐ Significance level	<p>The probability that a test statistic that is as extreme or more extreme than the actual value could have arisen by chance, given the null hypothesis.</p> <p>The lower the significance level, the less probable it is that the null hypothesis holds. In our case, the lower the significance level, the more likely it is that the two systems are significantly different.</p>

Table 1: Definition of Terms

The final key element in hypothesis testing is some means of generating the probability distribution of the test statistic under the assumption that the null hypothesis is true. Instead of assuming a distribution, such as the Normal distribution, we empirically generate the distribution as illustrated in Figure 1. As shown, the significance level is the area under the distribution curve bounded on the lower end by the actual value of the test statistic. The *significance level* is the probability that a test statistic that is as extreme or more extreme than the actual value could have arisen by chance, given the null hypothesis. Thus, the lower the significance level, the more likely it is that the two systems are significantly different.

Figure 1: Histogram of the Absolute Value of the Difference in Recall Scores

RANDOMIZATION TESTING

Traditional statistical analysis requires knowledge of or an assumption about the distribution of the data in the sample. We do not know the distribution for our sample and prefer to not make an assumption about it. Computationally-intensive methods empirically generate the distribution. Exact randomization testing generates all of the logically possible outcomes and compares the actual outcome to these. This amount of generation is often impractical because of the large number of data points involved, so approximate randomization is used instead. A confidence level is calculated to indicate how close the approximate randomization is to the exact randomization.

The method of approximate randomization involves random shuffling of the data to generate the distribution. We used the approximate randomization method described by Noreen in [1] with stratified shuffling to control for categorical variables that are not of primary interest in the hypothesis test.

The first step in approximate randomization is to select a test statistic. Our formulation of the null hypothesis indicates that the test statistic is the absolute value of the difference in recall, precision, or F-measure. The next step is to input the data. The data for each test set consists of the four-tuples of number possible, actual, correct, and partially correct for each message for each system. The actual statistic is calculated for the pair of systems as they come under scrutiny. The desired number of shuffles is set to 9,999 because it takes about eight hours to run the test for each test set and the confidence levels can be easily looked up in a table given in [1]. We have arbitrarily chosen the cutoff confidence level to be a conservative 99%. In a test of how the confidence levels were affected by the number of shuffles, 9,999 shuffles produced slightly higher confidence levels than 999 and were worth the 16-fold increase in computing time. Once the desired number of shuffles is set, the counters for the number of shuffles, ns , and the number of times the pseudostatistic is greater than or equal to the actual statistic, nge , are set to 0. A loop then increments ns until it has exceeded 9,999, the desired number of shuffles. The first step in this loop is to shuffle the data, which is the first major operation that occurs during approximate randomization. Table 2 contains an outline of the major operations involved in approximate randomization. Figure 2 illustrates the stratified shuffling used for the analysis of the MUC-4 results.

APPROXIMATE RANDOMIZATION	
1.	Shuffle ns times (ns is 9,999 in our case).
2.	Count the number of times (number greater than or equal, nge) that $ \text{stat}_{\text{pseudoA}} - \text{stat}_{\text{pseudoB}} \geq \text{stat}_A - \text{stat}_B $ (stat can be recall, precision, or F-measure in our case).
3.	The estimate of the significance level is $(nge + 1) / (ns + 1)$ (the 1's are added to ensure the test is valid).
4.	The confidence level is found by calculation or table lookup.

Table 2: Operations in Approximate Randomization

SHUFFLING

Figure 2: Shuffling for MUC-4

In Figure 2, data is shuffled by exchange of the systems' message scores depending on the outcome of a computer-simulated coin flip. After 100 coin flips, one per message, the absolute value of the difference in the metrics of the resulting pseudosystems can be compared to the corresponding absolute value of the difference in the metrics of the actual systems. The inequality for the comparison is shown in operation 2 of Table 2. The value of *nge* is incremented every time a randomized pair of pseudosystems satisfies this inequality. The significance level is calculated according to the formula in operation 3 of Table 2. The corresponding confidence level is then found in the appropriate table in [1].

According to Noreen,

Randomization is used to test the generic null hypothesis that one variable (or group of variables) is unrelated to another variable (or group of variables). Significance is assessed by shuffling one variable (or set of variables) relative to another variable (or set of variables). Shuffling ensures that there is in fact no relationship between the variables. If the variables are related, then the value of the test statistic for the original unshuffled data should be unusual relative to the values of the test statistic that are obtained after shuffling.¹

In our case, the four-tuple of data associated with each message for each system is the dependent set of variables that is shuffled. The explanatory variable is the system. Shuffling ensures that there is no relationship between the differences in the scores and the systems that produced them, i.e., that the differences were achieved by chance. If they were not achieved by chance, then the value of the actual test statistic will be far enough out on the tail of the empirically-generated distribution to be significant. The area under the distribution curve with the lower bound of the actual test statistic will be smaller than the cutoff in this case. We have arbitrarily chosen the cutoff to be 0.1 because we are able to distinguish reasonable groupings of systems at this level. At lower cutoffs, too many systems are not significantly different. The choice of a cutoff has traditionally been based on the practical implications of the choice. We need more data before we will know more of the implications of the cutoff given the sample sizes that we now have.

The stratified shuffling that was limited to exchange at the message level was the most conservative approach to shuffling because it eliminated the effect of the varying number of slots filled per message in the key. We did not want to introduce the effect of this nuisance variable. Instead, we wanted the simplest, most straightforward method to be used for the initial introduction of approximate randomization in this application. Further experimentation with the parameters of stratification, the cutoff for significance level, and the cutoff for confidence level may occur in the future.

Examples

These examples were obtained from [3] and are meant to give an intuitive sense of the results of the approximate randomization method. In general, if two systems have a large recall difference, they are likely to produce more homogeneous pseudo systems; *nge* will be small and the significance level will be low. On the other hand, if two systems have similar recall scores, they are likely to produce a high *nge* and a larger significance level, giving a lower likelihood that the differences are significant.

A pair of messages that show consistent differences across a range of messages will be more likely to be statistically significantly different in their score than two systems which have a small number of differences appearing in just a few messages. Consider the following systems reporting precision results for a set of 100 messages. System A has a score of 15/20 on each of 50 relevant messages and no spurious templates, for a total of $750/1000 = 75\%$ precision. System B has the identical score on each of the relevant messages except for one, where its score is 0/20. System B has a precision of $735/1000 = 73.5\%$. In the random shuffle, one of the pseudo systems will have the "0 precision" template. The pseudostatistic will always be the same as the measured absolute difference between the systems, i.e., 1.5%, so the significance level is 1.0 and the difference is not statistically significant.

1. From page 9 of [1].

Let us suppose that there is a System C with a precision score of 18/20 on each of 50 relevant messages and with no spurious templates. Any random shuffle of Systems A and C is likely to produce a smaller difference than the absolute value of the difference between A and C. The significance level will be extremely close to zero, indicating with virtual certainty that the two systems are significantly different.

WHAT WE CANNOT CONCLUDE FROM APPROXIMATE RANDOMIZATION

It is the systems themselves that set the conditions for the shuffle in the pairwise comparison. The same difference in scores between pairs of systems may not equate to the same significance level when different systems are involved in the pairwise comparison. The statistical test measures the difference from what would have occurred randomly. Each pair of systems provides a different set of conditions for the stratified shuffle. Thus, we cannot conclude from the results of the approximate randomization tests that the evaluation scores are generally significant within a certain range of percentage points. This cautionary note has further consequences which are discussed below in the results for TST3.

We also cannot conclude from the approximate randomization test whether our sample is sufficiently representative of newswire reports on activities concerning Latin America. In addition, we do not know whether the size of the sample is large enough for all of our purposes. The test sets were carefully collected in such a way as to reduce bias by selecting the number of articles concerning certain countries based on the percentage of articles about those countries in the entire corpus for the associated time period. In addition, the test articles were selected based solely on the order in which they appeared in the larger corpus and not on their contents. The size of the test sets was determined by the practical limits of time allotted for composing answer keys, running the participating systems on the test sets, and scoring the results. Even with these precautions, however, we still do not know whether the random sample of test messages is representative of the population from which they were drawn. Our statistical testing is not aimed at answering this question.

WHAT WE CAN CONCLUDE FROM APPROXIMATE RANDOMIZATION

TST3

The results for TST3 of MUC-4 are presented in this section. TST3 is the official test set for MUC-4. The systems officially report scores for recall, precision, and the F-measures for three weightings of recall and precision based on the ALL TEMPLATES row in the summary score report.² Approximate randomization was applied to each pair of participating systems. The approximate randomization method applied to the TST3 results randomly shuffles the message-by-message scores for the two systems being compared and checks to see how many times the test statistic is as extreme as or more extreme than the actual test statistic. The lower the number of times the pseudo test statistic is greater than or equal to the actual test statistic, the lower the significance level and the more likely the actual test statistic did not occur by chance.

The results of this computer-intensive testing are reported in two formats. The first format shows the significance levels for the pairwise comparison for each of the scores reported (Figures 3 - 6). The second and more informative format shows the significance groups or clusters based on the cutoffs at 0.1 for significance level and 0.99 for confidence level (Figures 7 - 11). The significance groupings represent groups of systems which were not considered to be significantly different from each other according to these cutoff criteria. Please note that the F-measures are calculated using floating point arithmetic throughout and differ slightly from the F-measures in the official score reports, which were calculated based on the integer values of recall and precision. These more accurate F-measures depicted in Figures 9 - 11 appear in tabular format in Appendix G.

2. For further information on the score report, see the paper "MUC-4 Evaluation Metrics" in these proceedings.

	BBN	CE	QECMU	HUGHES	LSI	NDC	MITRE	NM-BR	NYU	PARA	PRC	SRA	SRI	UM-CO	UMASS	UMICH	USC
BBN 30 44	0.0001 0.0128	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
CE 58 54	0.0002 0.0119	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
QECMU 49 55	0.0001 0.0112	0.0012 0.6144	0.0004 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
HUGHES 30 18	0.9582 0.0001	0.0001 0.0001	0.0004 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
LSI 23 16	0.0196 0.0001	0.0001 0.0001	0.0001 0.0001	0.0258 0.3082	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
NDC 20 30	0.0017 0.0007	0.0001 0.0001	0.0001 0.0001	0.0024 0.0001	0.3696 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
MITRE 12 8	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0103 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
NM-BR 22 25	0.0016 0.0001	0.0001 0.0001	0.0001 0.0001	0.0077 0.0026	0.6809 0.0004	0.5617 0.0807	0.0002 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
NYU 41 47	0.0133 0.458	0.0001 0.0359	0.0156 0.0189	0.0234 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
PARA 42 22	0.0027 0.0001	0.0011 0.0001	0.1343 0.0001	0.0027 0.0512	0.0001 0.002	0.0001 0.0008	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
PRC 28 41	0.4321 0.5441	0.0001 0.0014	0.0001 0.0017	0.4749 0.0001	0.1415 0.0001	0.0287 0.0037	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0002 0.0001	0.8215 0.0283	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
SRA 27 32	0.3748 0.0366	0.0001 0.0001	0.0001 0.0001	0.4444 0.0001	0.3218 0.0001	0.1199 0.5399	0.0001 0.0001	0.1278 0.0547	0.0009 0.0044	0.0001 0.0001	0.8215 0.0283	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
SRI 44 55	0.0003 0.0176	0.0003 0.8094	0.1611 0.9878	0.0008 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.4506 0.0522	0.5063 0.0001	0.0001 0.0001	0.0002 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
UM-CO 2 21	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.3876	0.0001 0.1122	0.0001 0.1098	0.0001 0.0001	0.0001 0.3389	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
UMASS 47 57	0.0001 0.0026	0.0009 0.3913	0.5281 0.5687	0.0002 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0559 0.0057	0.2496 0.0001	0.0001 0.0003	0.0001 0.0001	0.3246 0.5204	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001
UMICH 41 40	0.0022 0.2418	0.0001 0.0001	0.0223 0.0002	0.0084 0.0001	0.0001 0.0001	0.0001 0.0078	0.0001 0.0001	0.0001 0.0002	0.9559 0.0443	0.7194 0.0001	0.0001 0.7373	0.0004 0.0578	0.419 0.0001	0.0001 0.0005	0.0396 0.0001	0.0001 0.0001	0.0001 0.0001
USC 7 15	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.3493	0.0001 0.9413	0.0001 0.0018	0.0419 0.0018	0.0001 0.0042	0.0001 0.0001	0.0001 0.0185	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0176 0.346	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001

Legend:	System Recall, Precision	Recall Significance, Precision Significance	.05 ≤ significance level < .1 * .01 ≤ significance level < .05 ** significance level < .01 ***	(*) : conf < 0.99
---------	-----------------------------	--	---	-------------------

Figure 3: Significance Levels for Recall and Precision on TST3

	EBN	GE	GE-CHU	HUGHES	LSI	MDC	MITRE	NM-BR	NYU	PARA	PRC	SRA	SRI	UM-CO	UMASS	UMICH	USC
EBN		***	***	***	***	***	***	***	**	**		*	***	***	***		***
GE	0.0001		**	***	***	***	***	***	***	***	***	***	**	***	(*)	***	***
GE-CHU	0.0004	0.0415		***	***	***	***	***	***	***	***	***		***		***	***
HUGHES	0.0001	0.0001	0.0001				***		***	***(*)	***	*	***	***	***	***	***
LSI	0.0001	0.0001	0.0001	0.1208		**	***	**	***	***	***	***	***	***	***	***	***
MDC	0.0002	0.0001	0.0001	0.559	0.0314		***		***	***(*)	***(*)		***	***	***	***	***
MITRE	0.0001	0.0001	0.0001	0.0001	0.0002	0.0001		***	***	***	***	***	***	***	***	***	
NM-BR	0.0001	0.0001	0.0001	0.7759	0.0428	0.6827	0.0001		***	***	***	**	***	***	***	***	***
NYU	0.0389	0.0001	0.0053	0.0001	0.0001	0.0001	0.0001	0.0001		***	***	***		***	***		***
PARA	0.0117	0.0001	0.0001	0.0091	0.0001	0.0462	0.0001	0.0065	0.0001		*		***	***	***	***	***
PRC	0.3857	0.0001	0.0001	0.0019	0.0001	0.0077	0.0001	0.0007	0.0026	0.0914			***	***	***	**	***
SRA	0.0671	0.0001	0.0001	0.0538	0.0004	0.181	0.0001	0.0331	0.0001	0.9009	0.1882		***	***	***	***	***
SRI	0.0005	0.0285	0.325	0.0001	0.0001	0.0001	0.0001	0.0001	0.1451	0.0001	0.0001	0.0001		***		***	***
UM-CO	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0468	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		***	***	*
UMASS	0.0001	0.0994	0.8918	0.0001	0.0001	0.0001	0.0001	0.0001	0.0074	0.0001	0.0001	0.0001	0.3258	0.0001		***	***
UMICH	0.1681	0.0001	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	0.2613	0.0002	0.0131	0.0004	0.0048	0.0001	0.0001		***
USC	0.0001	0.0001	0.0001	0.0001	0.0007	0.0001	0.8106	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0593	0.0001	0.0001	

Legend:

F-Measure Significance
P&R equal weighting

(*) : conf < 0.99

* .05 ≤ significance level < .1
* .01 ≤ significance level < .05
* * significance level < .01

Figure 4: Significance Levels for the F-Measures with Recall and Precision Weighted Equally for TST3

	BBN	GE	GECHU	HUGHES	LSI	MDC	MITRE	NM-BR	NTU	PARA	PRC	SRA	SRI	UM-CQ	UMASS	UMICH	USC
BBN		***	***	***	***	***	***	***		***		**	***	***	***		***
GE	0.0004			***	***	***	***	***	***	***	***	***		***		***	***
GECHU	0.0018	0.5792		***	***	***	***	***	*** (*)	***	***	***		***		***	***
HUGHES	0.0001	0.0001	0.0001			***	***	*** (*)	***	**	***	***	***	***	***	***	***
LSI	0.0001	0.0001	0.0001	0.2282		***	***	***	***	***	***	***	***	***	***	***	
MDC	0.0003	0.0001	0.0001	0.0031	0.0002		***		***		***		***	***	***	***	***
MITRE	0.0001	0.0001	0.0001	0.0001	0.0002	0.0001		***	***	***	***	***	***		***	***	
NM-BR	0.0001	0.0001	0.0001	0.0455	0.0016	0.2103	0.0001		***		***	*	***	***	***	***	***
NTU	0.1553	0.0034	0.0083	0.0001	0.0001	0.0001	0.0001	0.0001		***	*	***	*	***	***	*	***
PARA	0.0001	0.0001	0.0001	0.0288	0.0005	0.1692	0.0001	0.9601	0.0001		***	***	***	***	***	***	***
PRC	0.4443	0.0001	0.0003	0.0001	0.0001	0.0041	0.0001	0.0002	0.0415	0.0001		*** (*)	***	***	***		***
SRA	0.0345	0.0001	0.0001	0.0005	0.0001	0.3808	0.0001	0.0343	0.0002	0.0049	0.048		***	***	***	*** (*)	***
SRI	0.0032	0.4684	0.6612	0.0001	0.0001	0.0001	0.0001	0.0001	0.0698	0.0001	0.0003	0.0001		***		***	***
UM-CQ	0.0001	0.0001	0.0001	0.0002	0.0057	0.0002	0.8951	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		***	***	
UMASS	0.0004	0.9871	0.7414	0.0001	0.0001	0.0001	0.0001	0.0001	0.0043	0.0001	0.0002	0.0001	0.4096	0.0001		***	***
UMICH	0.9093	0.0001	0.0002	0.0001	0.0001	0.0008	0.0001	0.0001	0.0856	0.0001	0.3673	0.0094	0.0004	0.0001	0.0001		***
USC	0.0001	0.0001	0.0001	0.0074	0.1144	0.0002	0.1612	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.3115	0.0001	0.0001	

Legend:

F-Measure Significance
(P twice weight of R)

* .05 ≤ significance level < .1
 ** .01 ≤ significance level < .05
 *** significance level < .01

(*) : conf < 0.99

Figure 5: Significance Levels for the F-Measures with Precision Weighted Twice Recall for TST3

	EBN	GE	GE-CHU	HUGHES	LSI	MDC	MITRE	NM-BR	NYU	PARA	PRC	SRA	SRI	UM-CO	UMASS	UMICH	USC
EBN				*	***	***	***	***	**				***	***	***	**	***
GE	0.0001		***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
GE-CHU	0.0002	0.0046		***	***	***	***	***	***(*)	***	***	***	***	***	***	***	***
HUGHES	0.0629	0.0001	0.0001		*		***		***	***			***	***	***	***	***
LSI	0.0002	0.0001	0.0001	0.0541			***		***	***	***	*	***	***	***	***	***
MDC	0.0007	0.0001	0.0001	0.1112	0.784		***		***	***	*		***	***	***	***	***
MITRE	0.0001	0.0001	0.0001	0.0001	0.0002	0.0005		***	***	***	***	***	***	***	***	***	
NM-BR	0.0002	0.0001	0.0001	0.1359	0.5674	0.7894	0.0001		***	***	***	*	***	***	***	***	***
NYU	0.0196	0.0001	0.0099	0.0004	0.0001	0.0001	0.0001	0.0001	*(*)	***	***	***	***	***	*		***
PARA	0.2717	0.0001	0.0001	0.0034	0.0001	0.0001	0.0001	0.0001	0.0474		*	***(*)	***	***	***	*	***
PRC	0.403	0.0001	0.0001	0.4651	0.0037	0.0182	0.0001	0.0072	0.0004	0.0248			***	***	***	***	***
SRA	0.2079	0.0001	0.0001	0.7917	0.045	0.1289	0.0001	0.0728	0.0003	0.0085	0.5577		***	***	***	***	***
SRI	0.0004	0.002	0.1954	0.0001	0.0001	0.0001	0.0001	0.0001	0.3091	0.0001	0.0001	0.0001		***			***
UM-CO	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0016	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		***	***	*
UMASS	0.0001	0.0042	0.6254	0.0001	0.0001	0.0001	0.0001	0.0001	0.026	0.0002	0.0001	0.0001	0.3194	0.0001		***	***
UMICH	0.0142	0.0001	0.0048	0.0001	0.0001	0.0001	0.0001	0.0001	0.6816	0.0675	0.0002	0.0002	0.1132	0.0001	0.0019		***
USC	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.1693	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.023	0.0001	0.0001	

(*): conf < 0.99

* .05 ≤ significance level < .1
 ** .01 ≤ significance level < .05
 *** significance level < .01

Legend:
 F-Measure Significance
 (P half weight of R)

Figure 6: Significance Levels for the F-Measures with Precision Weighted Half Recall for TST3

Figure 7: Significance Groupings for Recall at the 0.10 Level with 0.99 Confidence for TST3

Figure 8: Significance Groupings for Precision at the 0.10 Level with 0.99 Confidence for TST3

Figure 9: Significance Groupings for F-Measures (P&R) at the 0.10 Level with 0.99 Confidence for TST3

Figure 10: Significance Groupings for F-Measures (2P&R) at the 0.10 Level with 0.99 Confidence for TST3

Figure 11: Significance Groupings for F-Measures (P&2R) at the 0.10 Level with 0.99 Confidence for TST3

Although most of the information in the figures is self-explanatory, there are a few items of interest to be pointed out. The first concerns the kinds of point spreads that occur within and between significance groupings. For example, in Figure 7, the UMBC-Conquest team has a recall score of 2.5, which differs from the next higher recall score of 6.6 for USC by 4.1 percentage points. Whereas, the first large group of four systems, MDC (20.4), NMSU-Brandeis (21.9), LSI (23.0), and SRA (26.9), has a spread of 6.5 percentage points. This difference between what separates two significantly different systems and what separates systems which are not significantly different illustrates the cautionary note that this method is unable to determine that the evaluation scores are generally significant within a certain range of percentage points.

Another anomaly in the results is in the precision scores. In Figure 8, the UMBC-Conquest team has a precision score of 20.8. This score is not significantly different from seven other sites, although those seven sites divide into significance groupings of their own. The reason for this is the low number of actuals generated by the system.

Figure 9 illustrates two more examples of how the conditions set up by the systems influence the outcome of the statistical significance testing more than the actual test statistics themselves. The first example involves MDC, Paramax, and SRA and the second involves GE, GE-CMU, and UMass.

For the first example, the F-measures are as follows: MDC 24.33, Paramax 29.03, and SRA 29.33. Paramax and SRA are not significantly different at the 0.1 level with at least a confidence of 0.99; MDC and SRA are not, but MDC and Paramax are significantly different. If the cautionary note did not hold, then these results would be anomalous, because SRA's F-measure is higher than Paramax's. Therefore, we would expect that MDC and Paramax would not be significantly different because MDC and SRA are not. However, the F-measures are not the only factors affecting the results of the significance testing. The conditions set up by the systems during the shuffling have more influence than the raw results. This case illustrates that we cannot conclude from an approximate randomization test that the evaluation scores are generally significant within a certain range of percentage points.

In the case of GE, GE-CMU, and UMass, we see a similar illustration of the cautionary note. The F-measures

tures are: GE 56.01, GE-CMU 51.98, and UMASS 51.61. The significance levels are GE and GE-CMU: 0.0415, GE and UMASS: 0.0994, and GE-CMU and UMASS: 0.8918. GE and GE-CMU are significantly different at the 0.05 level and GE-CMU and UMASS are not significantly different even at the 0.1 level. The significance levels show that GE is significantly different from GE-CMU and UMASS at the 0.1 level. However, the confidence level for the significance test for GE and UMASS does not meet the 0.99 cutoff because it is only 0.635. So the pair cannot be considered significantly different by our criteria. The cautionary note explains why this situation could arise even though we would expect that the significance level would have been higher for GE and GE-CMU than for GE and UMASS.

TST4

The results for TST4 of MUC-4 are presented in this section. TST4 is a second test set chosen from an earlier time period. It contains more straightforward messages concerning terrorism with fewer irrelevant messages than TST3. The significance results are presented in the same formats as those for TST3. The first format is a matrix showing the significance levels for the pairwise comparison for each of the scores reported (Figures 12 - 15). The second and more informative format is a scatterplot showing the significance groups or clusters based on the cutoffs at 0.1 for significance level and 0.99 for confidence level (Figures 16 - 20). The significance groupings represent groups of systems which were not considered to be significantly different from each other according to these cutoff criteria.

SUMMARY

The results of the statistical analysis of the MUC-4 scores on TST3 and TST4 have been presented with an explanation of the computationally-intensive method of determining the statistical significance of the differences between systems. The method of approximate randomization was used to generate the probability distribution of the differences between systems. The statistical significance of the actual results can be ascertained by calculating the significance level during the generation of the distribution and looking up the confidence level in a published table. Knowing the statistical significance of the MUC-4 results allows us to draw valid conclusions concerning the performance of the participating systems.

ACKNOWLEDGEMENTS

I would like to thank Christopher Cooper for our discussions concerning the adaptation of approximate randomization to the MUC-4 problem and Lynette Hirschman and David Lewis for their help in explicating the statistical method for the computational linguistics community.

REFERENCES

- [1] Noreen, E. W. (1989) *Computer Intensive Methods for Testing Hypotheses: An Introduction*. New York: John Wiley & Sons.
- [2] Efron, B. and R. Tibshirani (1991) "Statistical Data Analysis in the Computer Age" *Science*, vol. 253, pp. 390 - 395.
- [3] Chinchor, N., L. Hirschman, and D. Lewis (1991) "Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3)" submitted to *Computational Linguistics*.

	BBN	GE	GE-CMU	HUGHES	LSI	MDC	MITRE	NM-BR	NYU	PARA	PRC	SRA	SRI	UN-CO	UMASS	UNICH	USC
BBN 30 44	...	...	...	...	...	...	...	...	...	...	...	..(*)	..	...	.	...	...
GE 54 54	0.0001 0.0299	...	...	...	...	...	...	...	...	..(*)	...	...	...	...	...	...	...
GE-CMU 49 55	0.0008 0.0321	0.003 0.8236	...	...	...	...	...	...	.	...	...	...	..	...	.	...	...
HUGHES 30 18	0.3139 0.0001	0.0001 0.0001	0.0006 0.0001	...	...	...	...	...	..	...	...	...	...	...	...	...	...
LSI 23 16	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0007 0.0049	...	...	...	...	...	...	...	...	...	...	...	...	...
MDC 26 30	0.0001 0.0022	0.0001 0.0001	0.0001 0.0001	0.0003 0.0001	0.812 0.0001	...	...	...	...	...	...	...	...	...	...	...	...
MITRE 12 8	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0004 0.0006	0.9122 0.4359	0.7322 0.0001	...	...	...	...	...	...	...	...	...	...	...
NM-BR 22 26	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0006 0.5829	0.6211 0.0005	0.7792 0.0002	0.4752 0.0001	...	...	...	...	...	...	...	...	...	...
NYU 41 47	0.1081 0.5278	0.0001 0.0921	0.1199 0.0771	0.023 0.0001	0.0001 0.0001	0.0001 0.0017	0.0001 0.0001	0.0001 0.0001	...	...	...	...	...	...	...	...	...
PARA 42 22	0.0214 0.0001	0.0097 0.0001	0.4332 0.0001	0.0004 0.4647	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.9575	0.454 0.0001	...	...	...	...	...	...	...	...
PRC 28 41	0.0014 0.4129	0.0001 0.0088	0.0001 0.0032	0.017 0.0001	0.0781 0.0001	0.1465 0.0326	0.0653 0.0001	0.2345 0.0001	0.0001 0.1972	0.0001 0.0001	...	..(*)	...	...	...	...	...
SRA 27 32	0.6422 0.009	0.0001 0.0001	0.0008 0.0001	0.6248 0.0001	0.0001 0.0001	0.0015 0.924	0.0001 0.0001	0.0015 0.0002	0.0668 0.0012	0.0071 0.0001	0.001 0.0497	...	...	...	...	...	...
SRI 44 55	0.1993 0.0416	0.0001 0.8476	0.0231 0.7058	0.0057 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.6625 0.174	0.2266 0.0001	0.0001 0.0057	0.1489 0.0001	...	...	..(*)	...	...
UN-CO 2 21	0.0001 0.0001	0.0001 0.0001	0.0001 0.0001	0.0001 0.5334	0.0001 0.0034	0.0001 0.0373	0.0001 0.0011	0.0001 0.9239	0.0001 0.0001	0.0001 0.9397	0.0001 0.004	0.0001 0.0725	0.0001 0.0001	...	...	...	...
UMASS 47 57	0.0613 0.7019	0.0005 0.0994	0.1962 0.0746	0.0004 0.0001	0.0001 0.0001	0.0001 0.0034	0.0001 0.0001	0.0001 0.0001	0.6775 0.7922	0.8763 0.0001	0.0001 0.2997	0.0207 0.0033	0.3573 0.0991	0.0001 0.0002	...	...	...
UNICH 41 40	0.3772 0.0065	0.0001 0.0001	0.0006 0.0001	0.835 0.0001	0.0008 0.0001	0.0002 0.4562	0.0006 0.0001	0.0026 0.0001	0.0547 0.0043	0.0064 0.0001	0.0051 0.0626	0.7112 0.4341	0.0375 0.0001	0.0001 0.0106	0.0051 0.0008	...	...
USC 7 15	0.0001 0.0064	0.0001 0.0004	0.0001 0.0002	0.0001 0.002	0.0127 0.0001	0.0025 0.4537	0.0156 0.0001	0.0011 0.017	0.0001 0.0029	0.0001 0.0065	0.0001 0.0355	0.0001 0.5755	0.0001 0.0001	0.0001 0.2288	0.0001 0.0015	0.0001 0.1939	...

Legend:

System Recall, Precision

Recall Significance, Precision Significance

* .05 ≤ significance level < .1
 ** .01 ≤ significance level < .05
 *** significance level < .01

(*) : conf < 0.99

Figure 12: Significance Levels for Recall and Precision on TST4

	BBN	GE	GE-CMU	HUGHES	LSI	MDC	MITRE	NL-BR	NYU	PARA	PRC	SRA	SRI	UN-CO	UMASS	UMICH	USC
BBN		***	***	***	***	***	***	***		***	***	**	*(*)	***		*(*)	***
GE	0.0001			***	***	***	***	***	***	***	***	***	**	***	***	***	***
GE-CMU	0.0009	0.1404		***	***	***	***	***	*	***	***	***	*	***	*	***	***
HUGHES	0.0001	0.0001	0.0001		***		***		***		***	***	***	***	***	***	**
LSI	0.0001	0.0001	0.0001	0.0019		***		**	***	***	***	***	***	***	***	***	
MDC	0.0001	0.0001	0.0001	0.4021	0.0004		***	(*)	***		*	*	***	***	***	***(*)	**
MITRE	0.0001	0.0001	0.0001	0.0002	0.5674	0.0001		***	***	***	***	***	***	***	***	***	
NL-BR	0.0001	0.0001	0.0001	0.3274	0.0153	0.0958	0.0013		***	*	***	***	***	***	***	***	
NYU	0.1612	0.0015	0.0626	0.0001	0.0001	0.0001	0.0001	0.0001		***	***	***		***		***	***
PARA	0.0002	0.0001	0.0001	0.1061	0.0002	0.5993	0.0001	0.0125	0.0001			**	***	***	***	*	***
PRC	0.007	0.0001	0.0001	0.0045	0.0001	0.0552	0.0001	0.0003	0.0001	0.1565			***	***	***		***
SRA	0.041	0.0001	0.0001	0.0034	0.0001	0.0536	0.0001	0.0001	0.0034	0.0241	0.6553		***	***	***		***
SRI	0.0468	0.0118	0.0907	0.0001	0.0001	0.0001	0.0001	0.0001	0.6164	0.0001	0.0001	0.0013		***		***	***
UN-CO	0.0001	0.0001	0.0001	0.0001	0.0026	0.0001	0.0111	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		***	***	***
UMASS	0.1742	0.0051	0.0731	0.0001	0.0001	0.0001	0.0001	0.0001	0.9618	0.0001	0.0002	0.0034	0.6521	0.0001		***	***
UMICH	0.0466	0.0001	0.0001	0.0001	0.0001	0.0083	0.0001	0.0001	0.0065	0.0254	0.3295	0.7873	0.0002	0.0001	0.0006		***
USC	0.0001	0.0001	0.0001	0.0422	0.3849	0.0127	0.1866	0.2759	0.0001	0.0032	0.0004	0.0005	0.0001	0.0003	0.0001	0.0001	

Legend:

F-Measure Significance
P&R equal weighting

* .05 ≤ significance level < .1
* .01 ≤ significance level < .05
* * * significance level < .01

(*) : conf < 0.99

Figure 13: Significance Levels for the F-Measures with Recall and Precision Weighted Equally for TST4

	BIB	CE	GE-CHU	HUGHES	LSI	MDC	MITRE	NM-BR	NYU	PARA	PRC	SRA	SRI	UM-CO	UMASS	UMICH	USC
BIB		***	**(*)	***	***	***	***	***		***	(*)	**	**	***		**	***
CE	0.0022			***	***	***	***	***	**	***	***	***		***	**	***	***
GE-CHU	0.0088	0.6794		***	***	***	***	***	*	***	***	***		***	*	***	***
HUGHES	0.0001	0.0001	0.0001		***	***	***		***		***	***	***	***	***	***	
LSI	0.0001	0.0001	0.0001	0.0029		***		***	***	***	***	***	***		***	***	***
MDC	0.0001	0.0001	0.0001	0.001	0.0001		***	***	***	**	*		***	***	***		*
MITRE	0.0001	0.0001	0.0001	0.0005	0.4739	0.0001		***	***	***	***	***	***		***	***	***
NM-BR	0.0001	0.0001	0.0001	0.9626	0.002	0.0018	0.0001		***		***	***	***	**	***	***	
NYU	0.3457	0.0231	0.0615	0.0001	0.0001	0.0001	0.0001	0.0001		***	*	***		***		***	***
PARA	0.0001	0.0001	0.0001	0.3124	0.0005	0.0196	0.0001	0.3754	0.0001		***	***	***	***	***	***	
PRC	0.0953	0.0001	0.0002	0.0001	0.0001	0.0308	0.0001	0.0001	0.0223	0.0001			***	***	**		***
SRA	0.014	0.0001	0.0001	0.0001	0.0001	0.4633	0.0001	0.0001	0.001	0.0001	0.2784		***	***	***		*
SRI	0.0348	0.2884	0.3575	0.0001	0.0001	0.0001	0.0001	0.0001	0.2778	0.0001	0.0001	0.0002		***		***	***
UM-CO	0.0001	0.0001	0.0001	0.0048	0.3708	0.0002	0.6293	0.0107	0.0001	0.0011	0.0001	0.0002	0.0001		***	***	***
UMASS	0.4395	0.04	0.0662	0.0001	0.0001	0.0001	0.0001	0.0001	0.8725	0.0001	0.0311	0.0029	0.2109	0.0001		***	***
UMICH	0.0111	0.0001	0.0001	0.0001	0.0001	0.1066	0.0001	0.0001	0.0035	0.0001	0.4593	0.5371	0.0001	0.0001	0.0004		***
USC	0.0002	0.0001	0.0001	0.3869	0.0005	0.0916	0.0001	0.46	0.0001	0.8753	0.0035	0.0531	0.0001	0.0072	0.0001	0.006	

Legend:

F-Measure Significance
(p twice weight of R)

* .05 ≤ significance level < .1
 ** .01 ≤ significance level < .05
 *** significance level < .01

(*) conf < 0.99

Figure 14: Significance Levels for the F-Measures with Precision Weighted Twice Recall for TST4

	BBN	GE	GE-CMU	HUGHES	LSI	MDC	MITRE	NIL-BR	NYU	PARA	PRC	SRA	SRI	UM-CO	UMASS	UMICH	USC
BBN		***	***	***	***	***	***	***			***			***	*		***
GE	0.0001		**	***	***	***	***	***	***	***	***	***	***	***	***	***	***
GE-CMU	0.0002	0.0147		***	***	***	***	***	*	***	***	***	*	***		***	***
HUGHES	0.0019	0.0001	0.0001		***	*(*)	***	**	***	***(*)			***	***	***	**	***
LSI	0.0001	0.0001	0.0001	0.0009		*			***	***	***	***	***	***	***	***	
MDC	0.0001	0.0001	0.0001	0.0492	0.0856		*(*)		***	***		***	***	***	***	***	***
MITRE	0.0001	0.0001	0.0001	0.0001	0.736	0.046		*	***	***	***	***	***	***	***	***	
NIL-BR	0.0001	0.0001	0.0001	0.0163	0.18	0.683	0.052		***	***	*	***	***	***	***	***	*
NYU	0.1069	0.0002	0.088	0.0001	0.0001	0.0001	0.0001	0.0001		*	***	*		***		*	***
PARA	0.4775	0.0001	0.0001	0.0078	0.0001	0.0013	0.0001	0.0001	0.0285		*		*	***	*		***
PRC	0.0017	0.0001	0.0001	0.8934	0.0009	0.1014	0.0002	0.0323	0.0001	0.031			*	***	***	*	***
SRA	0.2372	0.0001	0.0003	0.1144	0.0001	0.004	0.0001	0.0006	0.0158	0.6559	0.0617		*	***	***		***
SRI	0.1185	0.0001	0.0317	0.0001	0.0001	0.0001	0.0001	0.0001	0.9197	0.0328	0.0001	0.0229		***		***	***
UM-CO	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		***	***	***
UMASS	0.0751	0.0007	0.1217	0.0001	0.0001	0.0001	0.0001	0.0001	0.77	0.0107	0.0001	0.0062	0.6766	0.0001		***	***
UMICH	0.1911	0.0001	0.0001	0.0438	0.0001	0.0007	0.0001	0.0006	0.0247	0.6041	0.0258	0.8994	0.0075	0.0001	0.002		***
USC	0.0001	0.0001	0.0001	0.0001	0.2195	0.0041	0.3293	0.0125	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	

Legend:

F-Measure Significance
(P half weight of R)

* .05 ≤ significance level < .1
* .01 ≤ significance level < .05
* significance level < .01

(*) : conf < 0.99

Figure 15: Significance Levels for the F-Measures with Precision Weighted Half Recall for TST4

Figure16: Significance Groupings for Recall at the 0.10 Level with 0.99 Confidence for TST4

Figure 17: Significance Groupings for Precision at the 0.10 Level with 0.99 Confidence for TST4

Figure18: Significance Groupings for F-Measures (P&R) at the 0.10 Level with 0.99 Confidence for TST4

Figure 19: Significance Groupings for F-Measures (2P&R) at the 0.10 Level with 0.99 Confidence for TST4

Figure 20: Significance Groupings for F-Measures (P&2R) at the 0.10 Level with 0.99 Confidence for TST4