STATISTICAL SIGNIFICANCE OF MUC-6 RESULTS

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

INTRODUCTION

The results of the MUC-6 evaluation must be analyzed to determine whether close scores significantly distinguish systems or whether the differences in those scores are a matter of chance. In order to do such an analysis, a method of computer intensive hypothesis testing was developed by SAIC for the MUC-3 results and has been used for distinguishing MUC scores since that time. The implementation of this method for the MUC evaluations was first described in [1] and later the concepts behind the statistical model were explained in a more understandable manner in [2]. This paper gives the results of the statistical testing for the three MUC-6 tasks where a single metric could be associated with a system's performance.

STATISTICAL SIGNIFICANCE TESTING

Method

The general method employed to analyze the MUC-6 results is the Approximate Randomization method described in [3]. It is a computer intensive method which approximates the entire sample space in such a way as to allow us to determine the significance of the differences in F-Measures between each pair of systems and the confidence in that significance. The general method was applied on the basis of a message-by-message shuffling of a pair of MUC systems' responses to rule out differences that could have occurred by chance and to give us a picture of the similarities of the systems in terms of performance.

The method sorts systems into like and unlike categories. The results are shown in the following three tables for Named Entity, Template Element, and Scenario Template. These three all use the F-Measure as the single measure for systems as defined in [4] and in the MUC-6 Test Scores appendix to this proceedings. The parameters in the F-Measure used are such that recall and precision scores are combined with equal weighting. Note that Coreference was not characterized by F or any other unified measure because of the linkages that were being evaluated. Of course, an F-Measure is calculable, but more research is necessary before we can conclude that it will combine recall and precision in a way that is meaningful for these evaluations.

The statistical results reported here are based on the strictest cutoff point for significance level (0.01) and high confidence in the assigned level (at least 99%). What this method does not tell us is a numerical range within which F is not a significant distinguisher (such as plus or minus 3%). Instead it provides lists of similar systems. We have to be careful to not confuse the numerical order of the F-Measures with a ranking of systems and to instead look at the groupings on these charts. If a group or a single system is off by itself, then that group or single system is significantly different from its non-members. However, if there is overlap (and there is a lot of it in these results), then the ranking of the grouped systems is impossible. In addition, two similarly acting systems could use very different approaches to data extraction, so there may be some other value that distinguishes these systems that has not been measured in MUC-6.

Processing

To prevent human error, the entire process of doing the statistical analysis is automated. An awk program extracts tallies that appear in the score report output by the scoring software and puts them in a file to be fed to the C program for approximate randomization. The C program re-calculates F-measure, recall, and precision from raw tallies for higher accuracy than during the approximate randomization comparisons. The scoring program is slow in emacslisp and would be slowed further by calculations with higher accuracy. The statistical program outputs the significance and confidence levels in a matrix format for the analyst to inspect. Although 10,000 shuffles are carried out, the C program is fast. Results are depicted in lists of systems that are all equivalent, i.e., the differences in their scores were due to chance.

Results

The results are reported in a tabular format. The row headings contain the F-Measures for the systems and the rows are ordered from highest to lowest F. The columns are ordered in the same way as the rows and the headers contain the numerical order of the F values rather than the F value itself because of the size of the table on the page.

To use the table, you first determine which system you are interested in and identify its F-Measure in the left column, then look across the row or down the corresponding column to see which systems' F-Measures its F-Measure is not significantly different from. The systems that make up that group can be considered to have gotten their different F-Measures just by chance.

You can see, for instance, that among the Named Entity systems, the two lowest scoring systems are significantly different from each other and all of the all of the other systems. The two systems above them form a group which are significantly different from the other systems, but not from each other. A similar case appears in Template Element at the low and high end of the scores. However, the important thing to note is that there is a large amount of overlap otherwise. The Scenario Template test shows even more overlap than the other two tasks.

CONCLUSIONS

The groupings in these tables allow an ordering that is less clean than we would like, but that is realistic at this point in the evaluation methodology research. In addition to looking at the scores, evaluation research on a more granular level is needed to understand the differences in the systems' performance. Such research could reveal strengths and weaknesses in extracting certain information and lead to test designs that focus research in areas that will directly impact operational value. Also, other factors that are of interest to consumers, such as speed, development data requirements, and so on, need to be considered when making comprehensive comparisons of systems.

The entire community would benefit from more refined measured values and a better understanding of how the differences in human performance influence the results. Distinguishing systems at such a strict cutoff as we use in the statistics may only be justified if variations in human performance are smaller. After all, it is the human interpretation of the task definitions that informs the systems during development. Especially in Named Entity where machine performance and human performance are close, we would expect to see inherent human differences in interpreting language during both system and answer key development to be a considerable factor holding the machines back.

REFERENCES

[1] Chinchor, N., Hirschman, L., and D. Lewis (1993) "Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3) " Computational Linguistics 19(3).

[2] Chinchor, N. (1992). "The Statistical Significance of the MUC-4 Results" Proceedings of the Fourth Message Understanding Conference (MUC-4). Morgan Kaufmann, Publishers. San Mateo, CA.

[3] Noreen, W. (1989) Computer Intensive Methods for Testing Hypotheses: An Introduction. John Wiley & Sons.

[4] Van Rijsbergen, C.J. (1979) *Information Retrieval.* London: Butterworths.

NE Statistical Results

	Similar Systems																			
F	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
96.42	✓	✓	✓																	
95.66	✓	✓	✓	✓		✓														
94.92	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓									
94.00		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓									
93.65			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓								
93.33		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓								
92.88			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓								
92.74			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓								
92.61			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓							
91.20				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓							
90.84			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
89.06					✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
88.19									✓	✓	✓	✓	✓	✓	✓	✓				
85.82											✓	✓	✓	✓	✓	✓				
85.73											✓	✓	✓	✓	✓	✓				
84.95											✓	✓	✓	✓	✓	✓				
67.62																	✓	✓		
59.38																	✓	✓		
35.46																			✓	
2.38																				✓

TE Statistical Results

	Similar Systems											
F	1	2	3	4	5	6	7	8	9	10	11	12
79.99	✓	✓										
79.85	✓	✓		✓	✓							
77.31			✓	✓	✓	✓						
77.24		✓	✓	✓	✓	✓						
76.29		✓	✓	✓	✓	✓	✓	✓				
74.96			✓	✓	✓	✓	✓	✓				
74.32					✓	✓	✓	✓				
71.97					✓	✓	✓	✓	✓	✓		
71.16							✓	✓	✓	✓		
69.80								✓	✓	✓		
61.17											✓	
53.80												✓

ST Statistical Results

	Similar Systems										
F	1	2	3	4	5	6	7	8	9	10	11
56.40	✓	✓	✓	✓	✓	✓		✓			
54.39	✓	✓	✓	✓	✓	✓	✓	✓			
53.27	✓	✓	✓	✓	✓	✓	✓	✓			
51.63	✓	✓	✓	✓	✓	✓	✓	✓	✓		
50.98	✓	✓	✓	✓	✓	✓	✓	✓	✓		
50.96	✓	✓	✓	✓	✓	✓	✓	✓	✓		
48.96		✓	✓	✓	✓	✓	✓	✓	✓		
48.14	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
43.24				✓	✓	✓	✓	✓	✓	✓	
40.35								✓	✓	✓	✓
33.44										✓	✓