THE STATISTICAL SIGNIFICANCE OF THE MUC-5 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 statistical significance of the results of the MUC-5 evaluation is determined using a computer-intensive method of hypothesis testing known as approximate randomization. The exact method is described in detail in [1] and [2] and has been used as the accepted statistical test for the MUC results since MUC-3. The purpose of the statistical testing is to determine whether the scores of the systems are different by chance or due to a significant difference in the character of the systems.

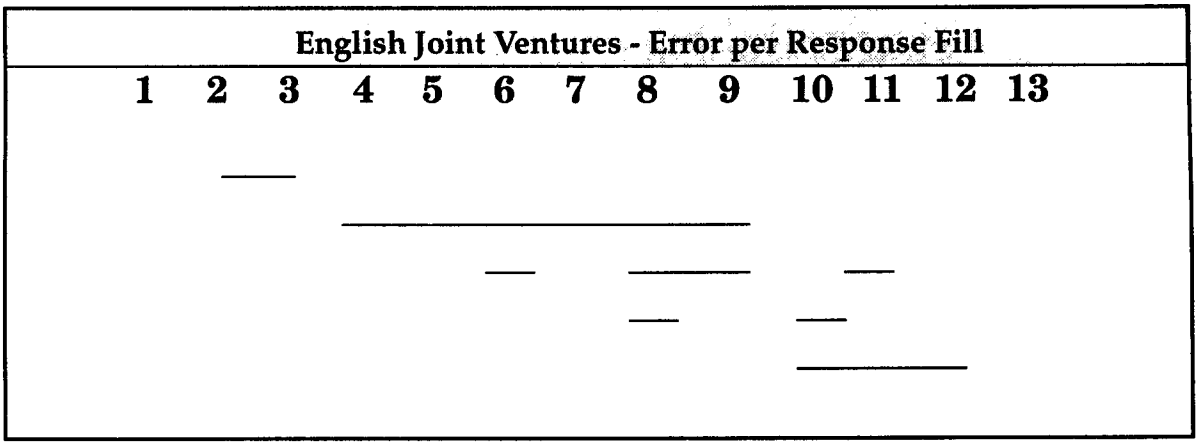
STATISTICAL SIGNIFICANCE RESULTS

Statistical significance results are reported here for the following metrics: Error per Response Fill, F-Measure with recall and precision weighted equally, and Richness-Normalized Error (minimum and maximum). The systems are compared for the same domain and language and, thus, there are four figures for each metric: English Joint Ventures (EJV), Japanese Joint Ventures (JJV), English Microelectronics (EME), and Japanese Microelectronics (JME). The format of the reporting is according to the groupings of the systems which are not significantly different from each other at the 0.01 level with a confidence of at least 99%. Systems which are not significantly different from each other are underscored on the same line. The systems are numbered to save space and the correspondence of the number and system site are given below the significance results.

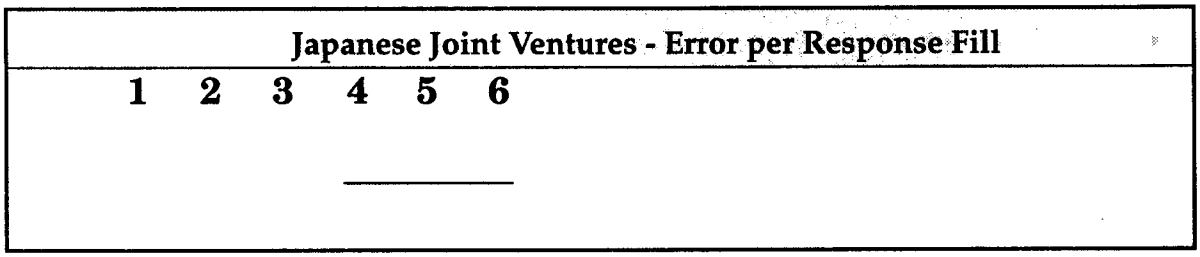
It is interesting to note that the rankings of systems do not change when using the Error per Response Fill metric or the F-Measure. The numerical rankings change slightly (numbers 6 and 7 in EJV reverse, and numbers 4 and 5 in JJV reverse), but those changes are not significant statistically because the two members in each of the reversed pairs are both in the same significance grouping for both of the two metrics. It is also interesting to note that the Error per Response Fill metric distinguishes four more systems than the F-Measure over all domains and languages. The Richness-Normalized Error metric distinguishes far fewer systems statistically than the Error per Response Fill metric with 29 systems distinguished by Richness-Normalized Error as opposed to 55 by Error per Response Fill for EJV alone. Both the minimum and maximum Richness-Normalized Error metrics produce the same rankings and statistical results so are conflated here. The statistical groupings of systems for Richness-Normalized Error are so large and so numerous that systems cannot be distinguished well enough to reflect their perceived differences in performance. It is believed that this is due to the fact that the Richness-Normalized Error metric ignores the amount of spurious data generated by a system and that the amount and kind of spurious data generated impacts the perception of how well the system is doing in an operational setting.

CONCLUSIONS

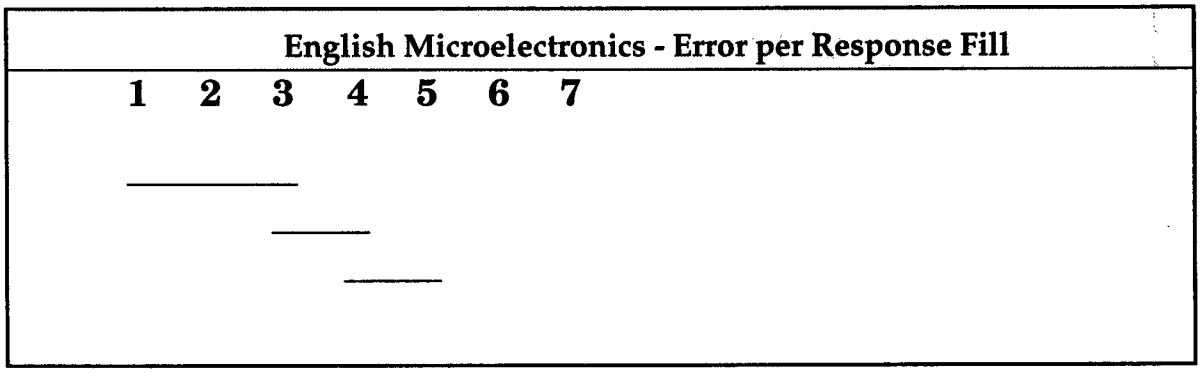
The approximate randomization method has been used to determine the statistical significance of the rankings of systems for MUC-5. It is also useful for reflecting on the relative merits of the evaluation metrics. The statistical results show that the Error per Response Fill metric is the most sensitive metric of the three in terms of distinguishing systems. However, no statistically significant changes in ranking occur when F-Measure is used. The Richness-Normalized Error metric distinguishes far fewer systems than either of the other metrics.



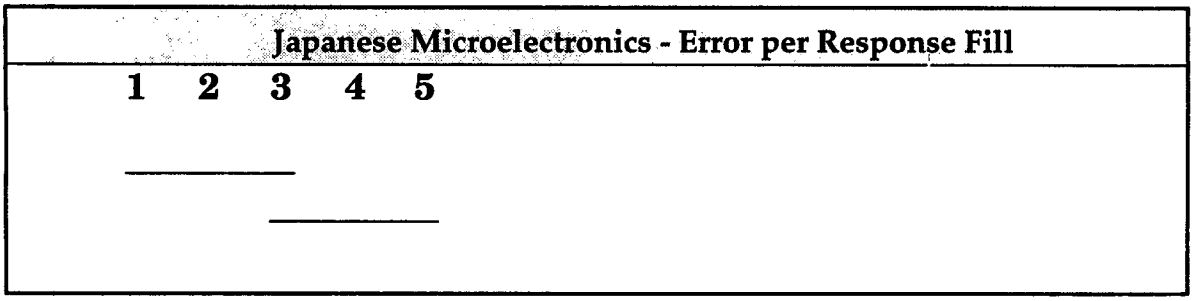
1) GE/CMU 2) BBN 3) SRI 4) UMASS/HU 5) PMAX 6) USUSSEX 7) NMSU/BR
8) NYU 9) SRA 10) PRC 11) USC 12) MITRE 13)TRW



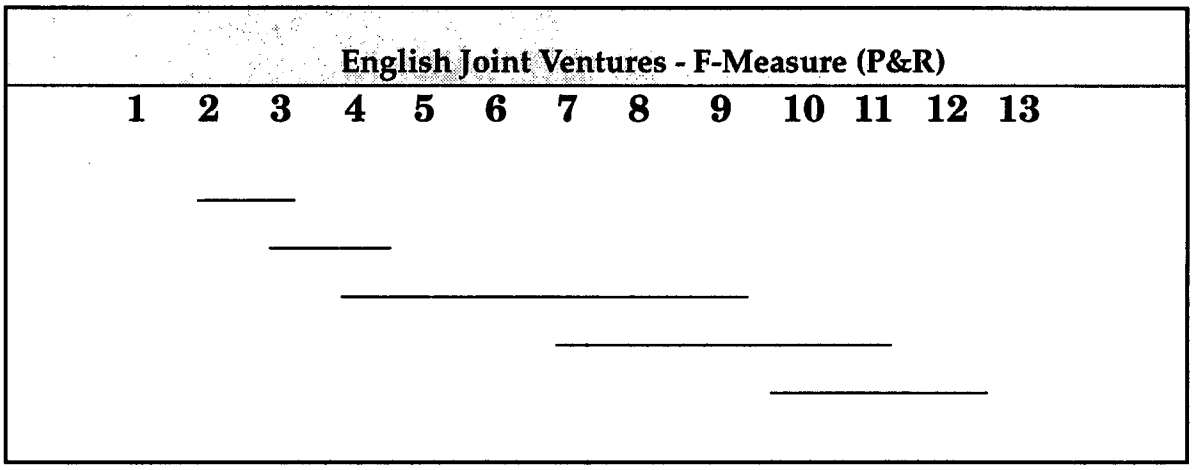
1) GE/CMU-OPT 2) GE/CMU 3) NMSU/BR 4) SRA 5) SRI 6) BBN



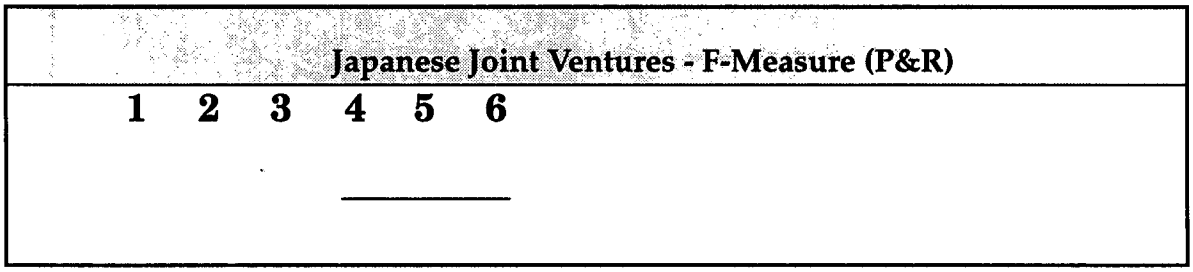
1) GE/CMU 2) BBN 3) UMAN 4) LSI 5) NMSU/BR 6) UMASS/HU 7) UMICH



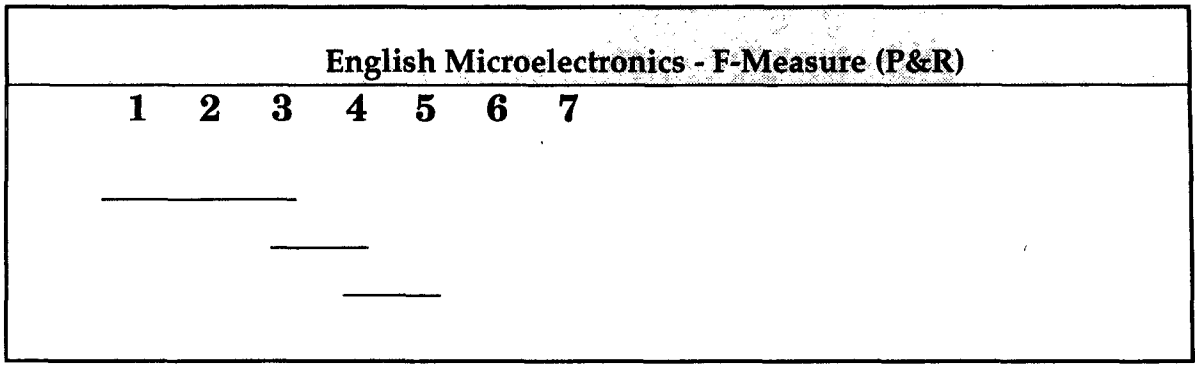
1) GE/CMU 2) GE/CMU-OPT 3) NMSU/BR 4) NEC 5) BBN



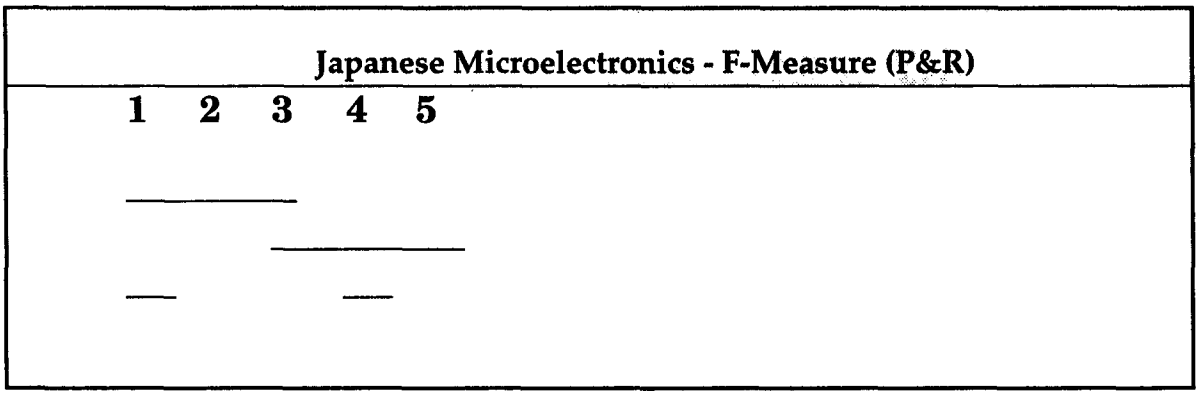
1) GE/CMU 2) BBN 3) SRI 4) UMASS/HU 5) PMAX 6) NMSU/BR 7) USUSSEX
8) NYU 9) SRA 10) PRC 11) USC 12) MITRE 13) TRW



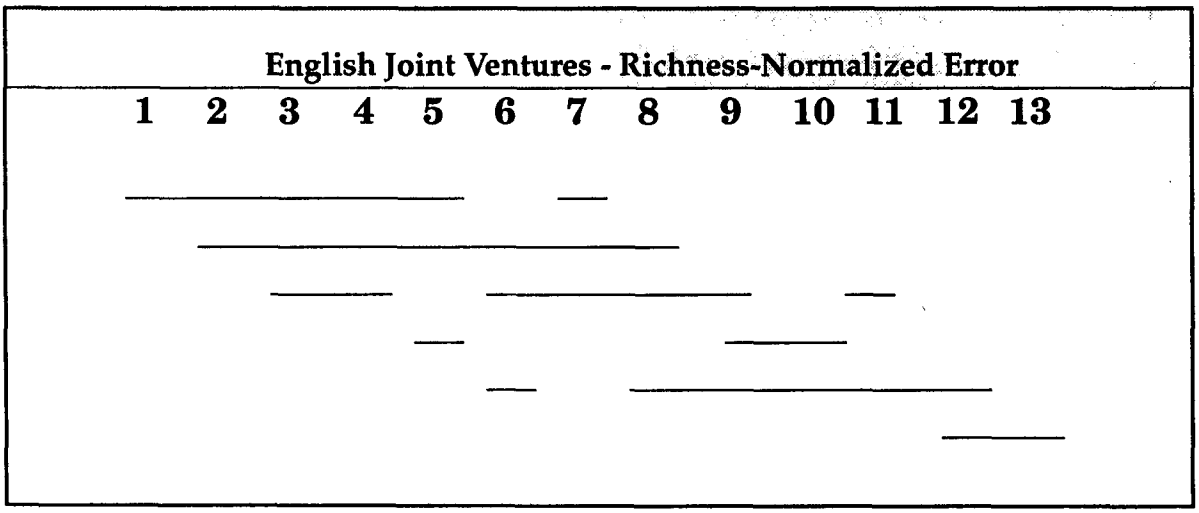
1) GE/CMU-OPT 2) GE/CMU 3) NMSU/BR 4) SRI 5) SRA 6) BBN



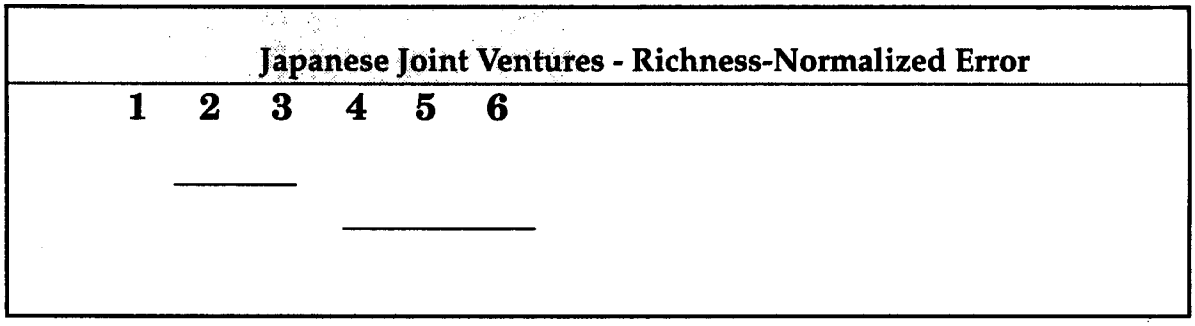
1) GE/CMU 2) BBN 3) UMAN 4) LSI 5) NMSU/BR 6) UMASS/HU 7) UMICH



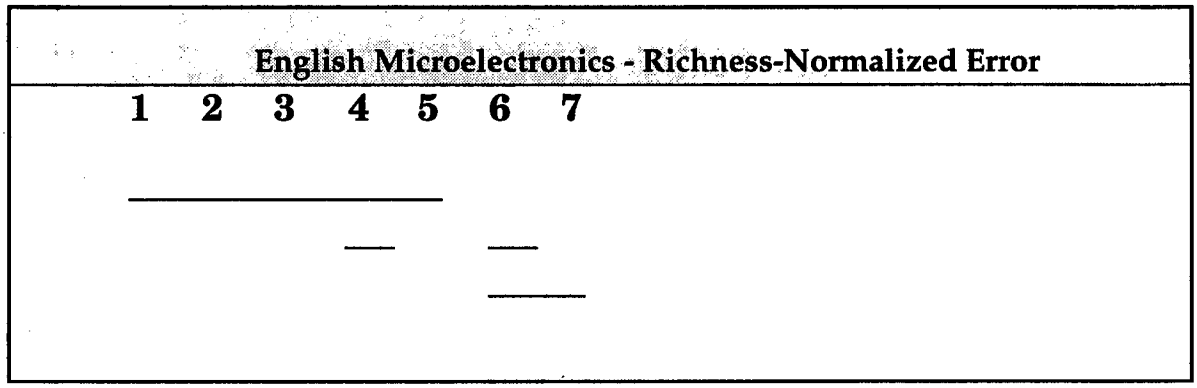
1) GE/CMU 2) GE/CMU-OPT 3) NMSU/BR 4) NEC 5) BBN



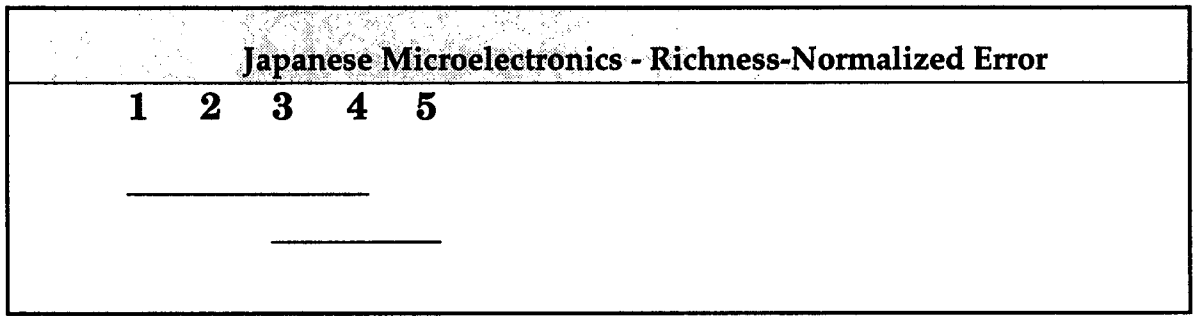
1) BBN 2) SRI 3) NYU 4) UMASS/HU 5) GE/CMU 6) USUSSEX 7) NMSU/BR
8) SRA 9) MITRE 10) PRC 11) PMAX 12) USC 13) TRW



1) GE/CMU-OPT 2) GE/CMU 3) NMSU/BR 4) SRI 5) BBN 6) SRA



1) BBN 2) LSI 3) NMSU/BR 4) UMAN 5) GE/CMU 6) UMICH 7) UMASS/HU



1) GE/CMU-OPT 2) NMSU/BR 3) NEC 4) GE/CMU 5) BBN

[1] Chinchor, N., L. Hirschman, 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.