APPENDIX H:

SCATTER PLOTS OF FINAL TEST RESULTS

This appendix contains selected scatter plots of much of the data found in appendix G. The data points in the scatter plots are labeled with abbreviations of the organization names, as follows:

BBN (BBN)
GE (GE)
GE-CM (GE and CMU)
HU (Hughes)
LSI (LSI)
MDC (MDESC)
MITR (MITRE)
NM-BR (NMSU and Brandeis)
NYU (NYU)
PMAX (Paramax)
PRC (PRC)
SRA (SRA)
SRI (SRI)
UM-CQ (UMBC and ConQuest)
UMA (UMass)
UMI (UMichigan)
USC (USC)

Most of the scatter plots are taken from the All Templates Row in the summary score report. The All Templates scores include severe penalties for missing and spurious data. See the paper on evaluation metrics by N. Chinchor in Part I for further information.

SECTION 1. OVERALL RESULTS FOR TST3 AND TST4

The plots in this section are taken from some of the rows that appear at the bottom of the summary score reports. Some of the data is discussed in the paper by B. Sundheim in Part I of this proceedings.

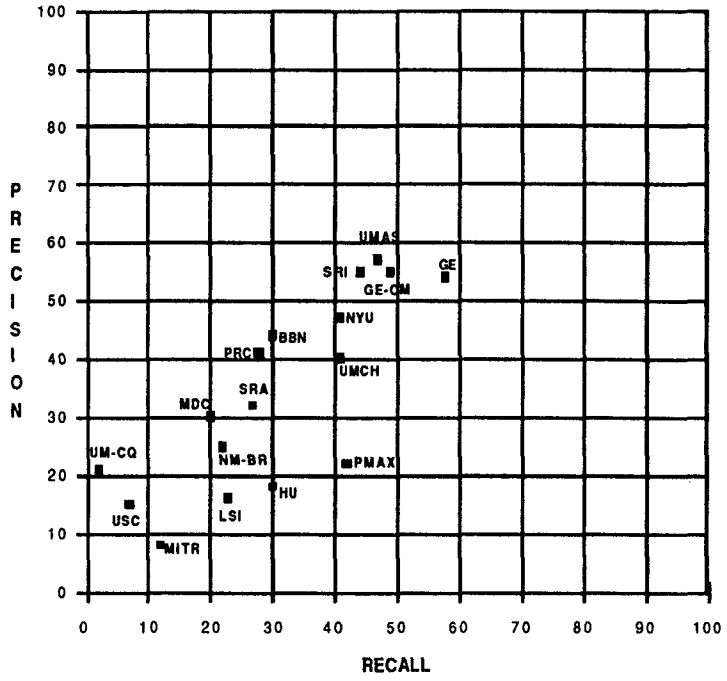

Figure H1. TST3 Base Run (Recall vs Precision): All Templates Row

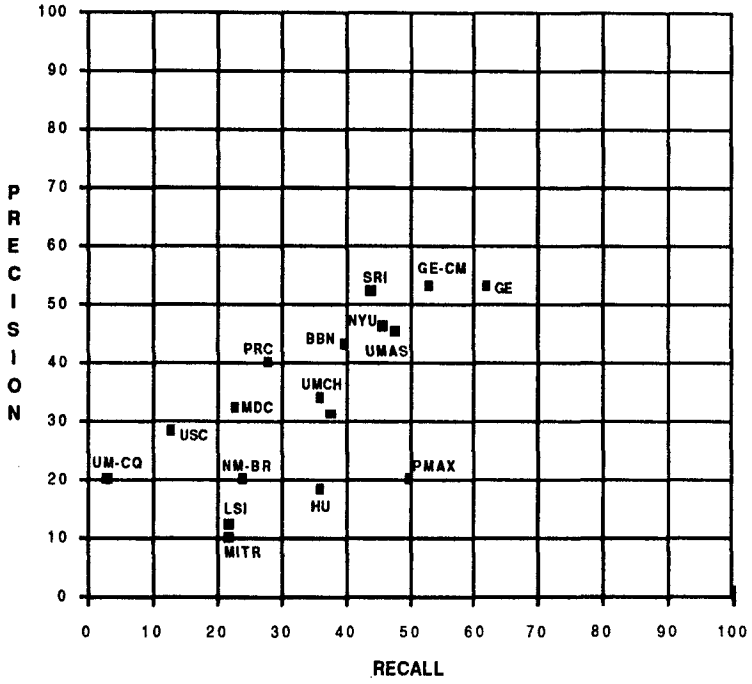

Figure H2. TST4 Base Run (Recall vs Precision): All Templates Row

Figures H3 and H4 plot data points only on systems for which the results of both optional test runs and base test runs were submitted. The labels on the data points give an indication of the type of test that was run. See section 3 in appendix G and the papers in Parts II and III for further information.

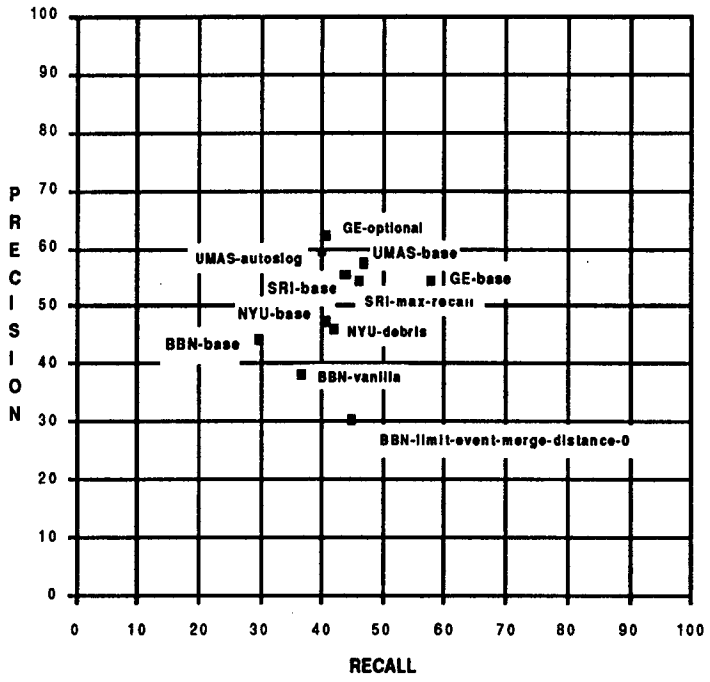

Figure H3. TST3 Base and Optional Runs (Recall vs Precision): All Templates Row

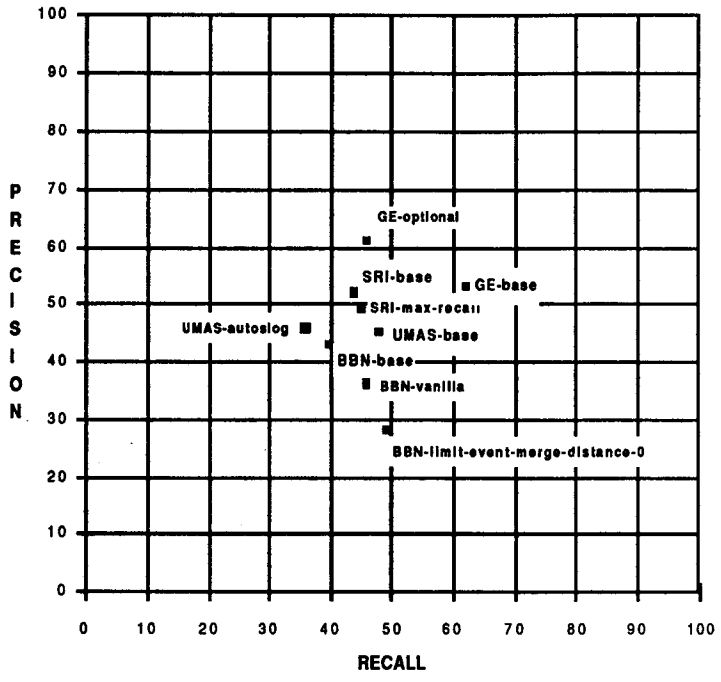

Figure H4. TST4 Base and Optional Runs (Recall vs Precision): All Templates Row

Figures H5 and H6 show how much spurious data (templates and slot values) was generated for the two test sets.

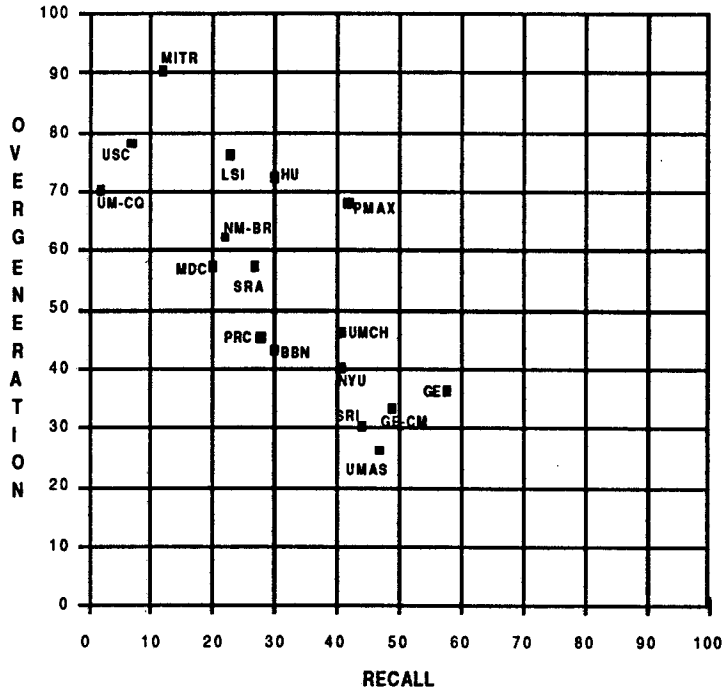

Figure H5. TST3 Base Run (Recall vs Overgeneration): All Templates Row

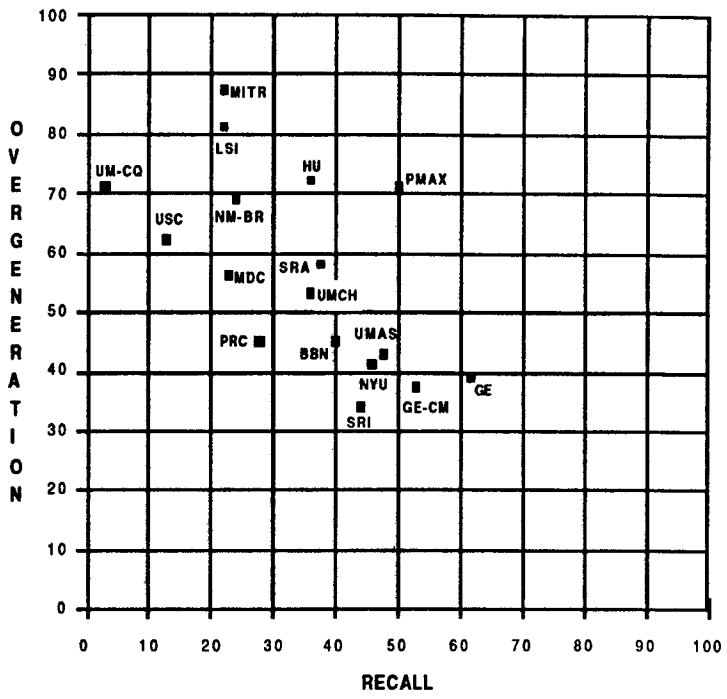

Figure H6. TST4 Base Run (Recall vs Overgeneration): All Templates Row

Figures H7, H8, and H9 compare the results of TST3 with those of TST4. Squares indicate data points for TST3; triangles indicate data points for TST4.

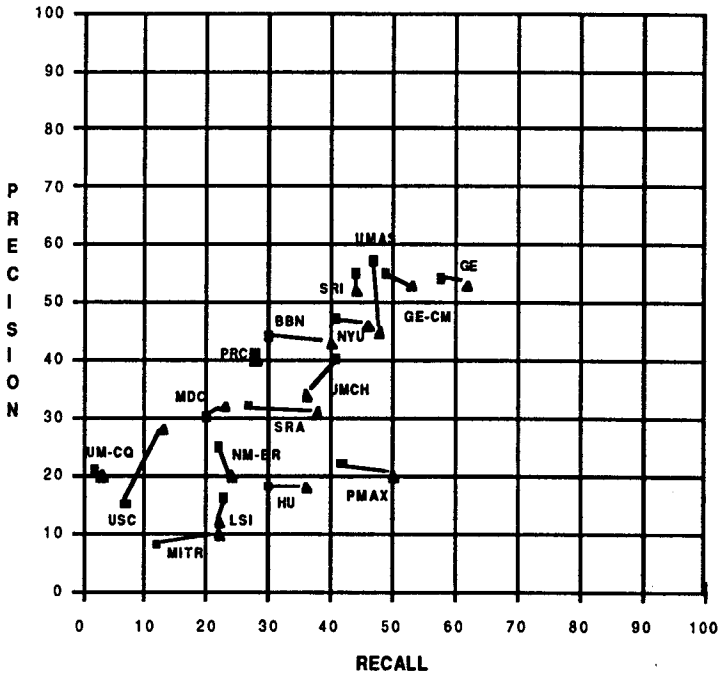

Figure H7. TST3 and TST4 Base Runs (Recall vs Precision): All Templates Row

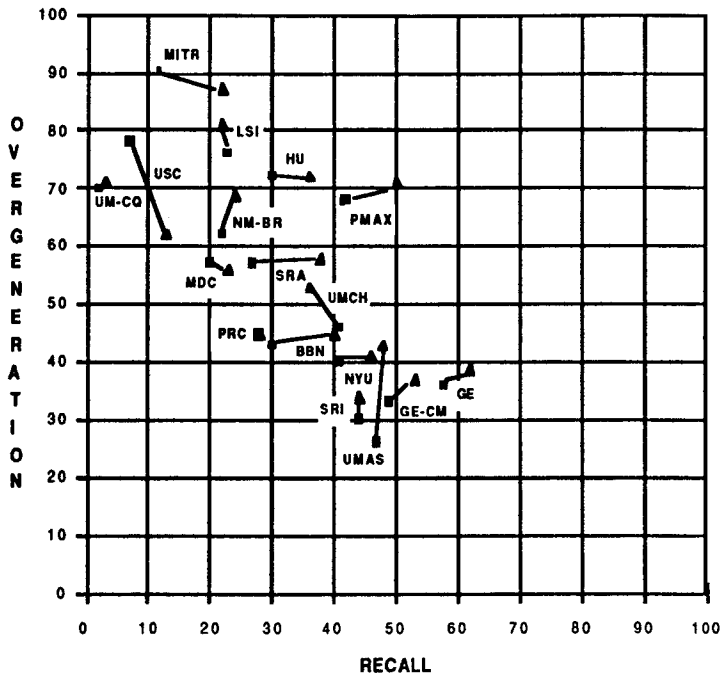

Figure H8. TST3 and TST4 Base Runs (Recall vs Overgeneration): All Templates Row

Figures H10, H11, and H12 show each system's "region of performance," which is defined by the scores in the Matched Only, Matched/Missing, Matched/Spurious, and All Templates rows. The Matched Only scores include only mild penalties for missing and spurious data; the Matched/Missing scores include severe penalties for missing data and mild penalties for spurious data; the Matched/Spurious scores include mild penalties for missing data and severe penalties for spurious data; the All Templates scores include severe penalties for both missing and spurious data. The systems have been divided into three groups in order to make the scatter plots less crowded.

Key: Upper left-hand corner = Matched/Missing Upper right-hand corner = Matched Only
Lower left-hand corner = All Templates Lower right-hand corner = Matched/Spurious

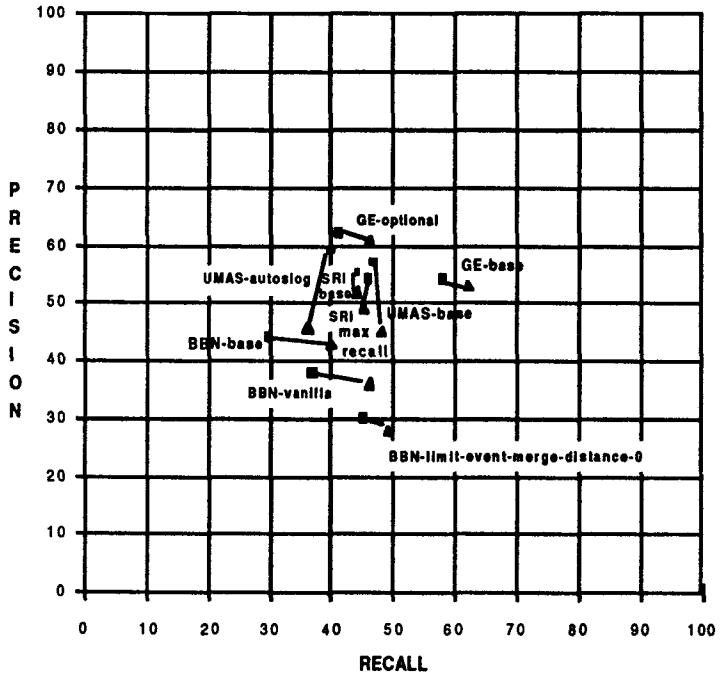

Figure H9. TST3 and TST4 Base and Optional Runs (Recall vs Precision): All Templates Row

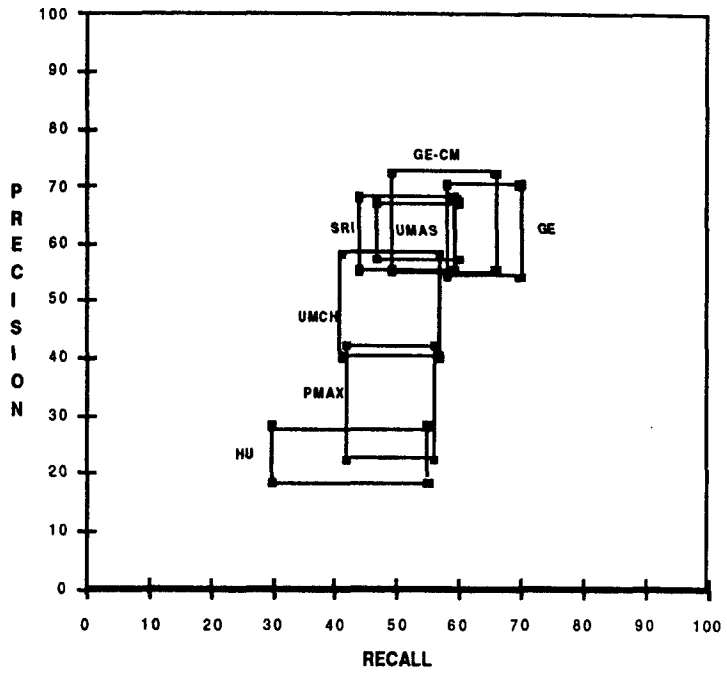

Figure H10. TST3 Base Run (Recall vs Precision): Smallest Regions of Performance

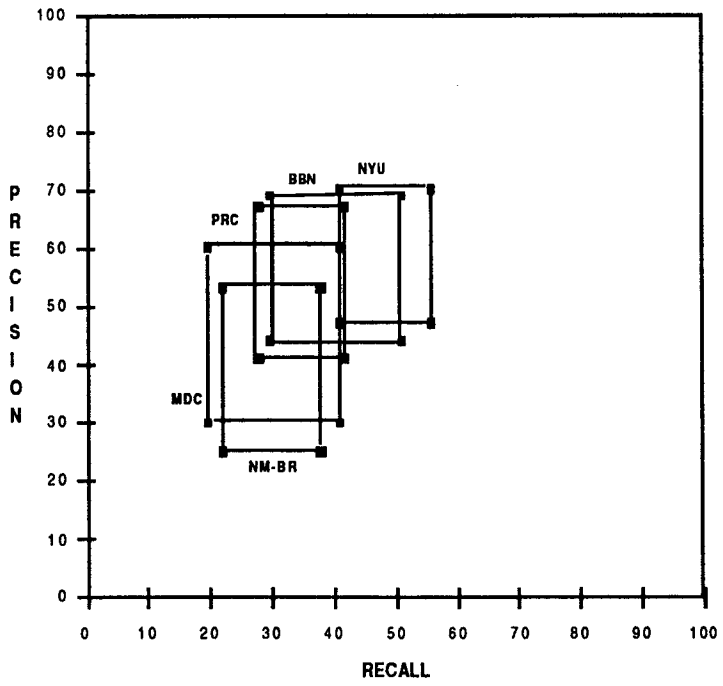

Figure H11. TST3 Base Run (Recall vs Precision): Larger Regions of Performance

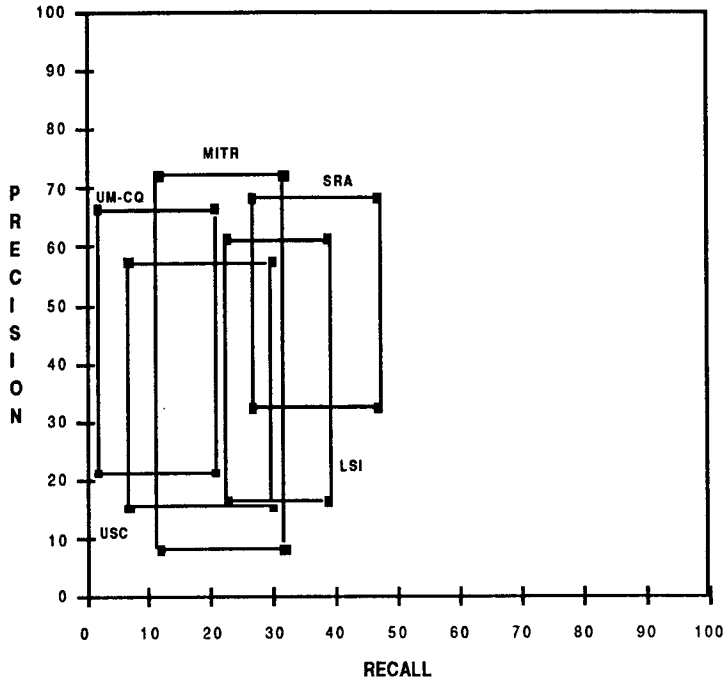

Figure H12. TST3 Base Run (Recall vs Precision): Largest Regions of Performance

Figures H13 and H14 show how well the systems were able to discriminate between relevant texts and irrelevant texts. See the paper by D. Lewis and R. Tong in Part I of this proceedings for further information.

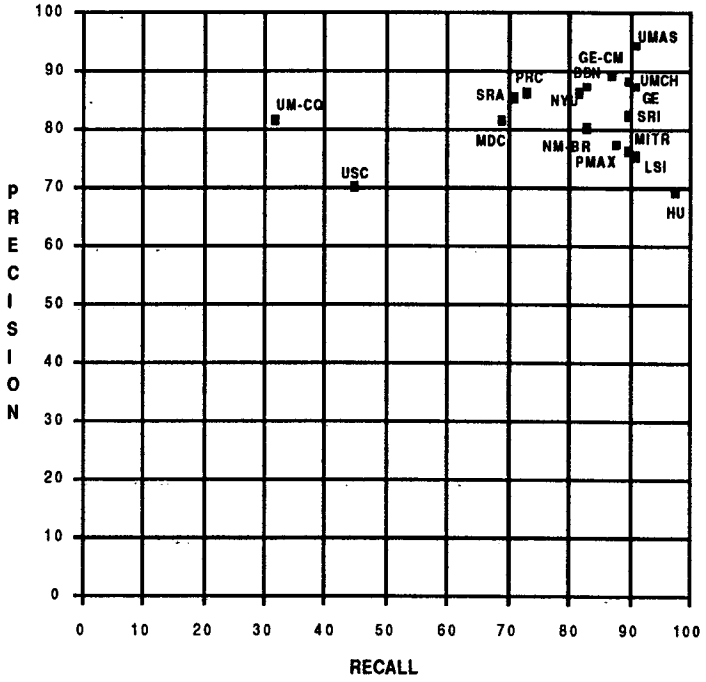

Figure H13. TST3 Base Run (Recall vs Precision): Text Filtering Row

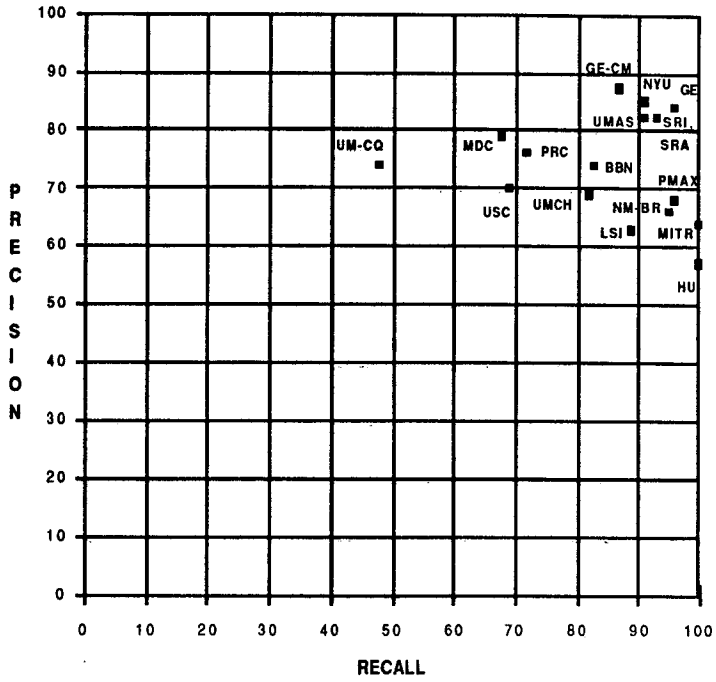

Figure H14. TST4 Base Run (Recall vs Precision): Text Filtering Row

SECTION 2. SUBTOTAL RESULTS FOR TST3 AND TST4

The plots in this section are taken from various places in the summary score reports. Some of the data is discussed in the paper by B. Sundheim in Part I of this proceedings.

Figures H15 and H16 plot the scores for those slots that require set fills (i.e., fills from a set of predefined alternatives).

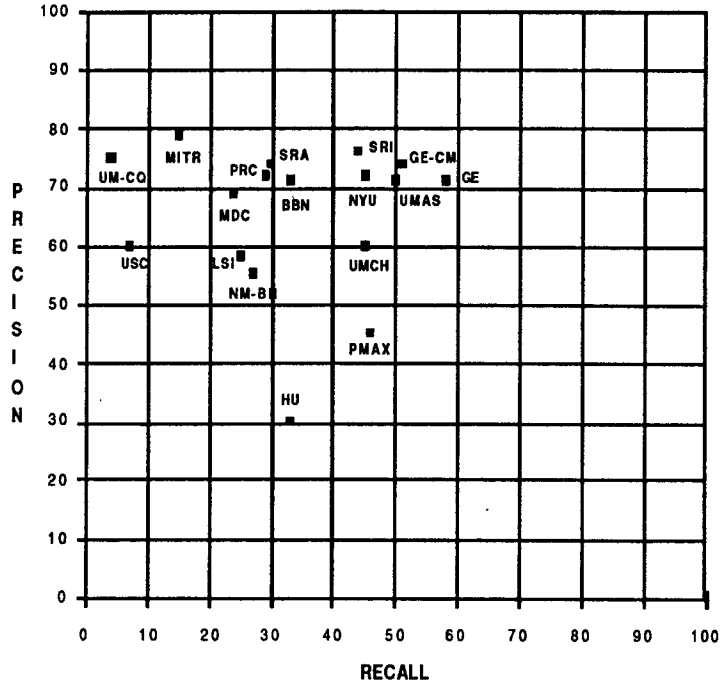

Figure H15. TST3 Base Run (Recall vs Precision): Set Fills Only

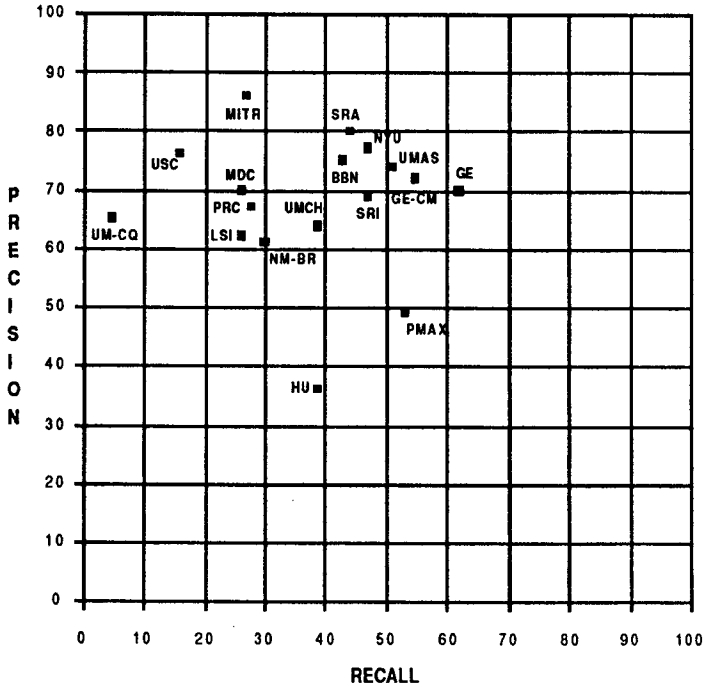

Figure H16. TST4 Base Run (Recall vs Precision): Set Fills Only

Figures H17 and H18 plot the scores for those slots that require string fills (i.e., fills that are unaltered text strings).

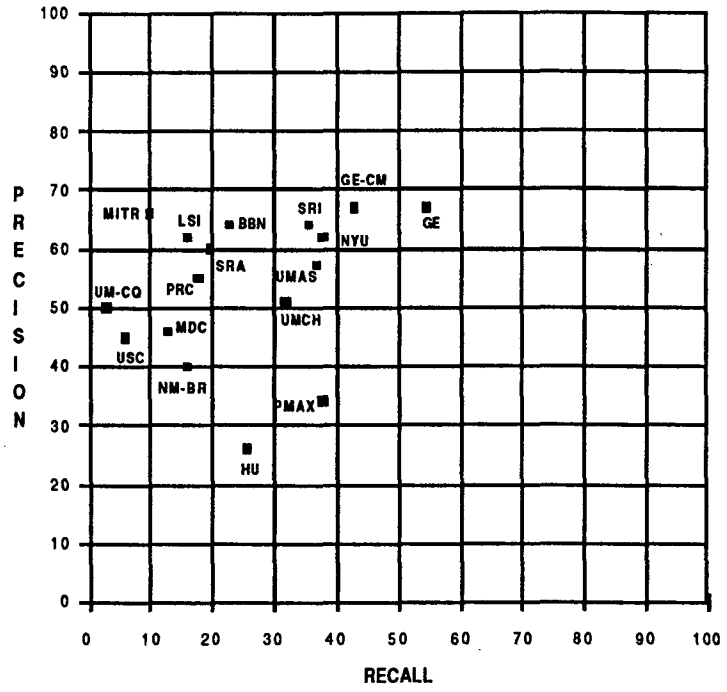

Figure H17. TST3 Base Run (Recall vs Precision): String Fills Only

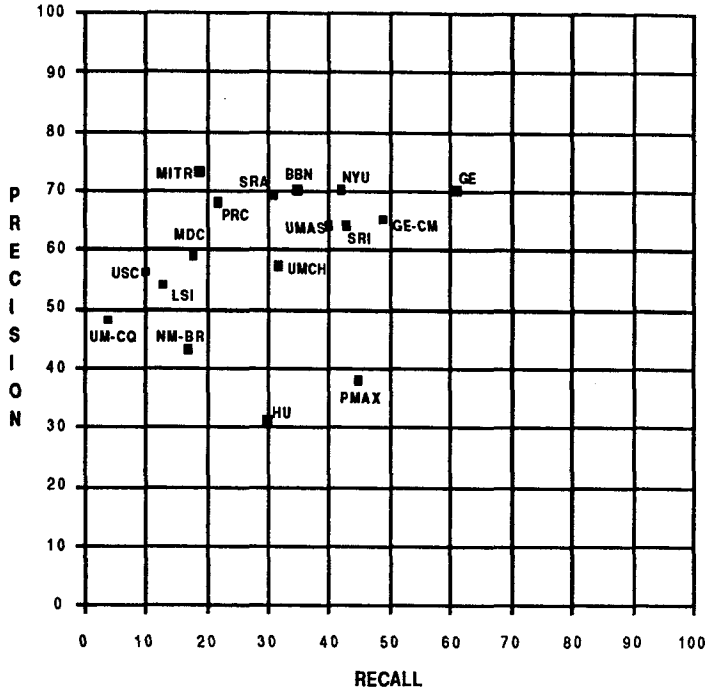

Figure H18. TST4 Base Run (Recall vs Precision): String Fills Only

Figures H19 and H20 plot the scores for the slots containing information about the incident (date, location, type, stage of execution, and instrument).

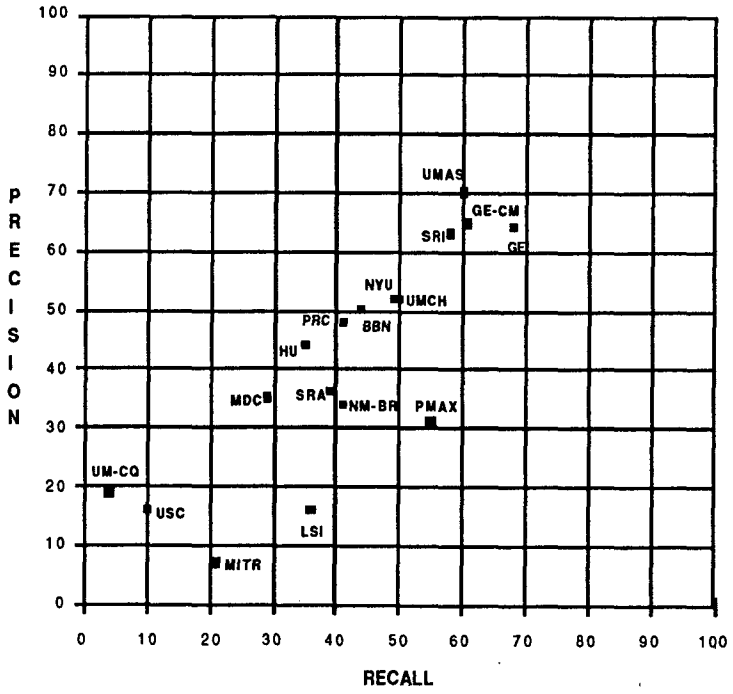

Figure H19. TST3 Base Run (Recall vs Precision): Incident Pseudo-Object

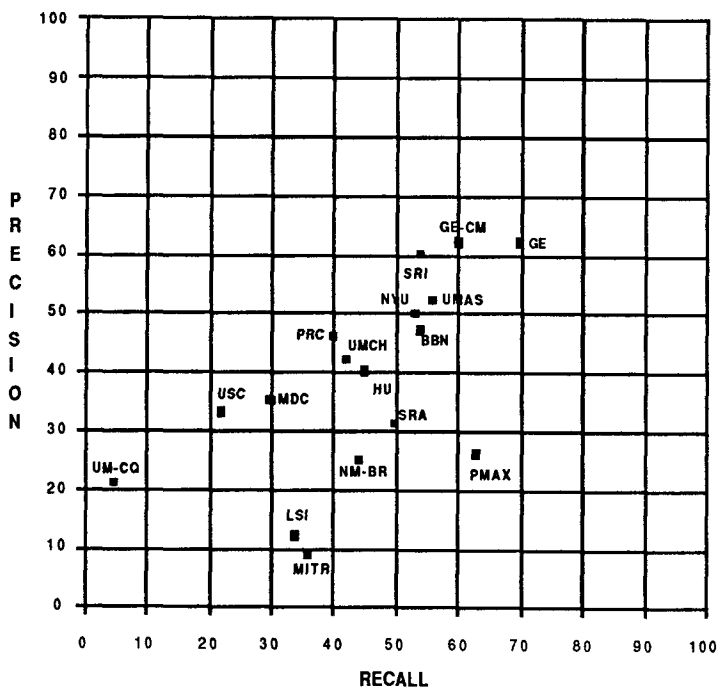

Figure H20. TST4 Base Run (Recall vs Precision): Incident Pseudo-Object

Figures H21 and H22 plot the scores for the slots containing information about the perpetrator (category, individual id, organization id, and organization confidence).

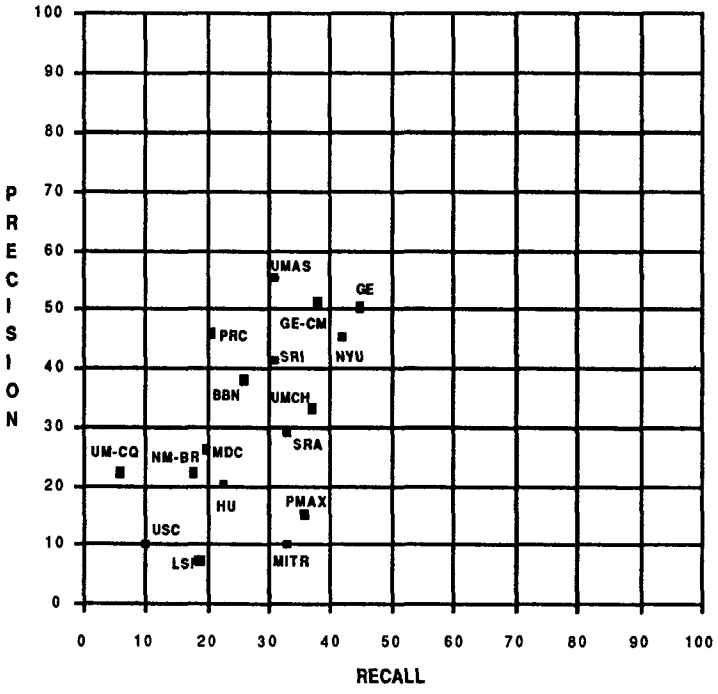

Figure H21. TST3 Base Run (Recall vs Precision): Perpetrator Pseudo-Object

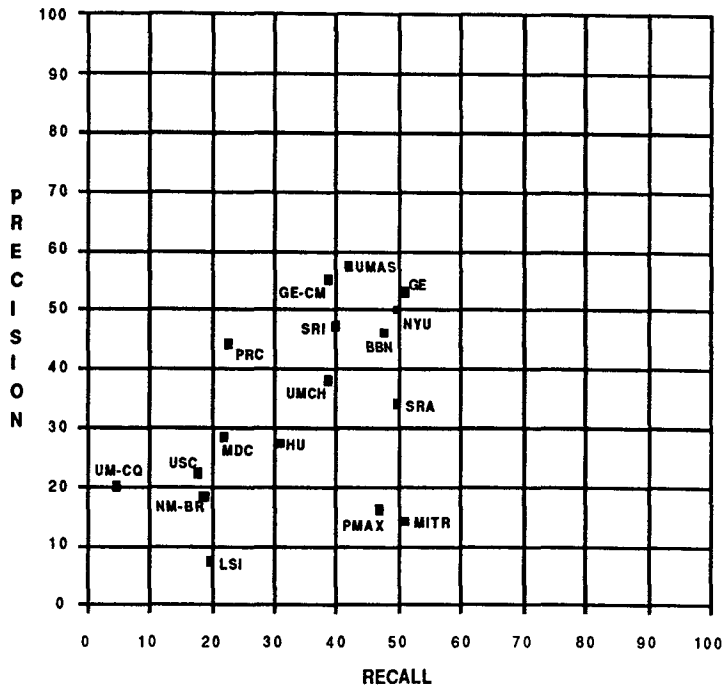

Figure H22. TST4 Base Run (Recall vs Precision): Perpetrator Pseudo-Object

Figures H23 and H24 plot the scores for the slots containing information about the physical target (id, type, number, foreign nationality, effect, and total number).

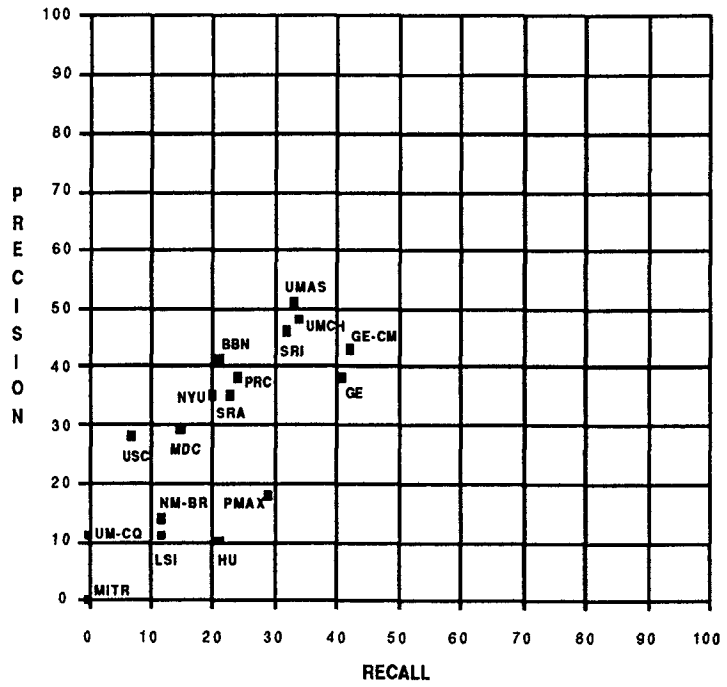

Figure H23. TST3 Base Run (Recall vs Precision): Physical Target Pseudo-Object

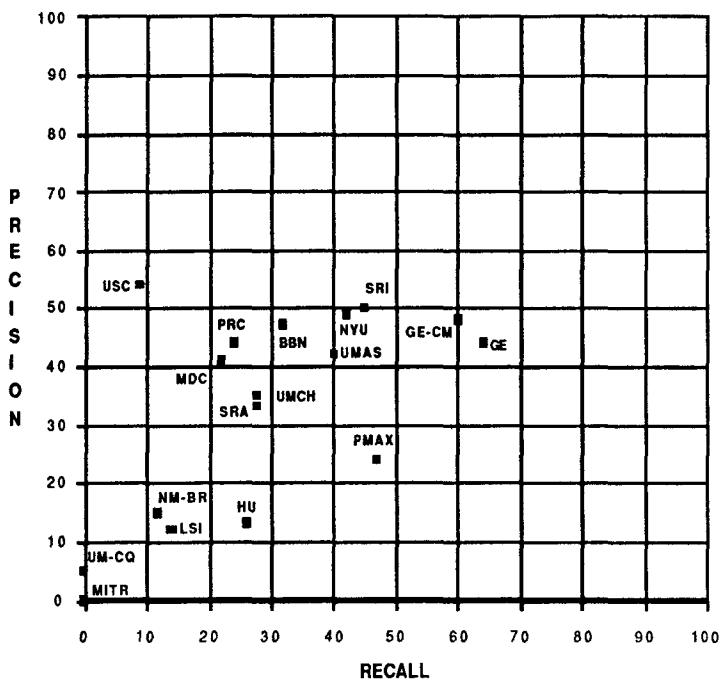

Figure H24. TST4 Base Run (Recall vs Precision): Physical Target Pseudo-Object

Figures H25 and H26 plot the scores for the slots containing information about the human target (name, description, type, number, foreign nationality, effect, and total number).

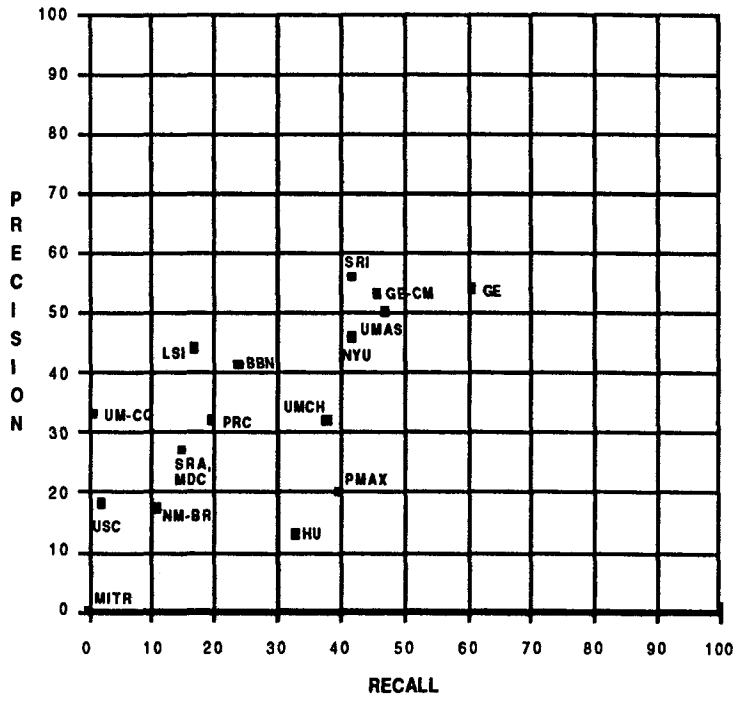

Figure H25. TST3 Base Run (Recall vs Precision): Human Target Pseudo-Object

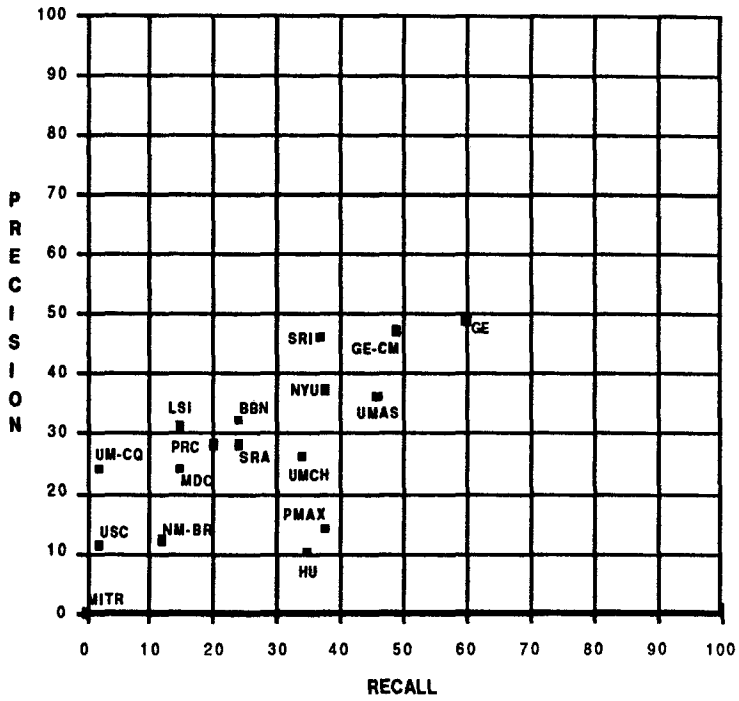

Figure H26. TST4 Base Run (Recall vs Precision): Human Target Pseudo-Object

Figures H27-H32 plot the scores for a few of the individual slots in the template, as a means of showing the way spurious data generation combines with incorrect data generation to affect precision. The only penalty to precision in the MESSAGE: TEMPLATE (template-id) slot is due to spurious data generation, since the template ID is an arbitrary number that is never scored incorrect.

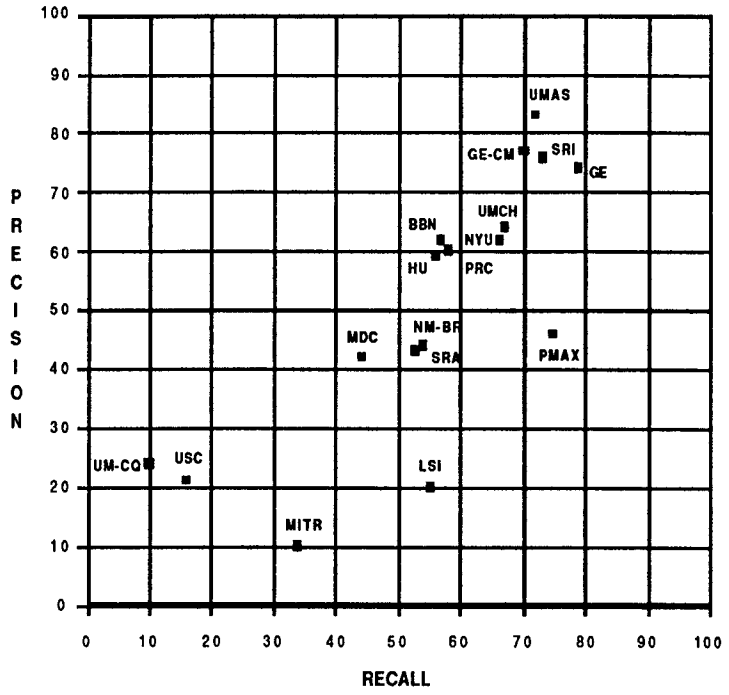

Figure H27. TST3 Base Run (Recall vs Precision): Slot 1 (template-id)

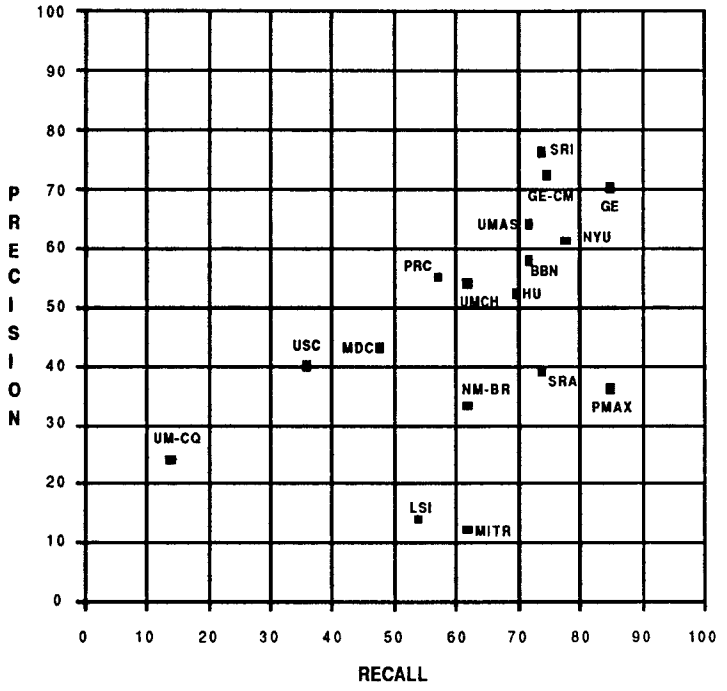

Figure H28. TST4 Base Run (Recall vs Precision): Slot 1 (template-id)

Since the INCIDENT: TYPE (inc-type) slot is single-valued and can never be null, its filler will never be scored spurious. Thus, the only penalty to precision in the inc-type slot is due to incorrect data generation rather than to spurious data generation.

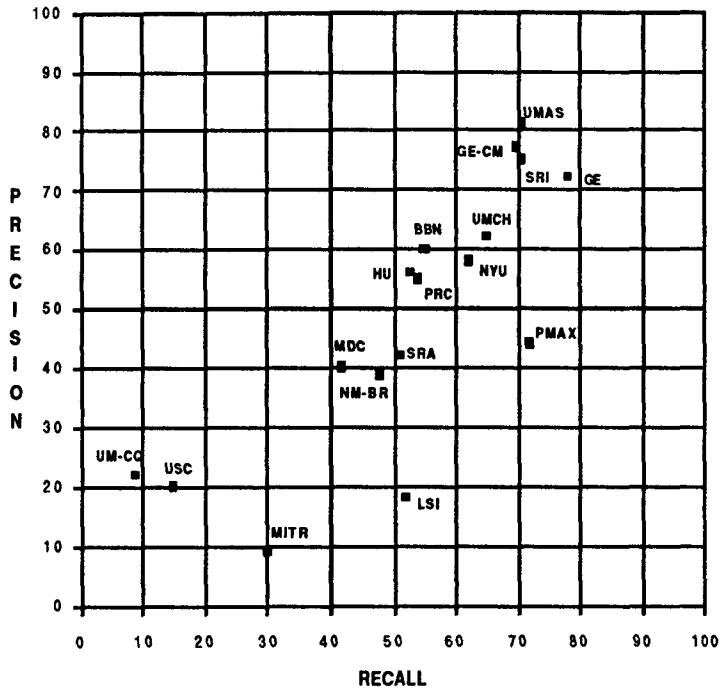

Figure H29. TST3 Base Run (Recall vs Precision): Slot 4 (inc-type)

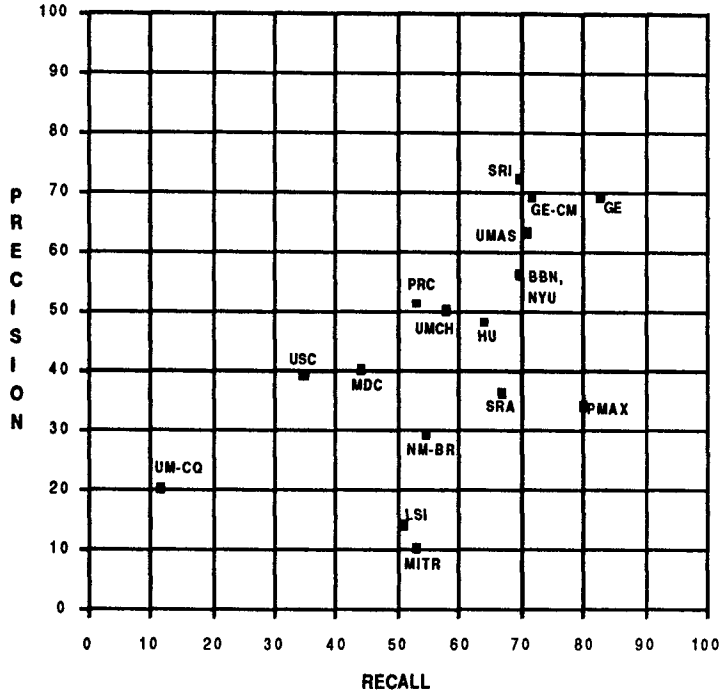

Figure H30. TST4 Base Run (Recall vs Precision): Slot 4 (inc-type)

Both spurious and incorrect data generation act as penalties on precision in the PHYSICAL TARGET: ID (phys-tgt-id) slot, which is a multi-valued slot that can be null, can contain a single value, or can contain an indefinite number of values.

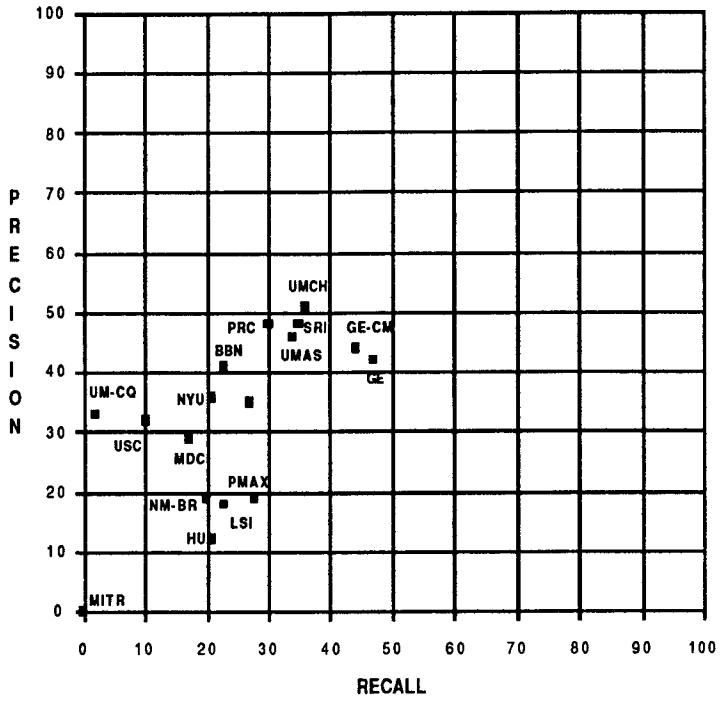

Figure H31. TST3 Base Run (Recall vs Precision): Slot 12 (phys-tgt-id)

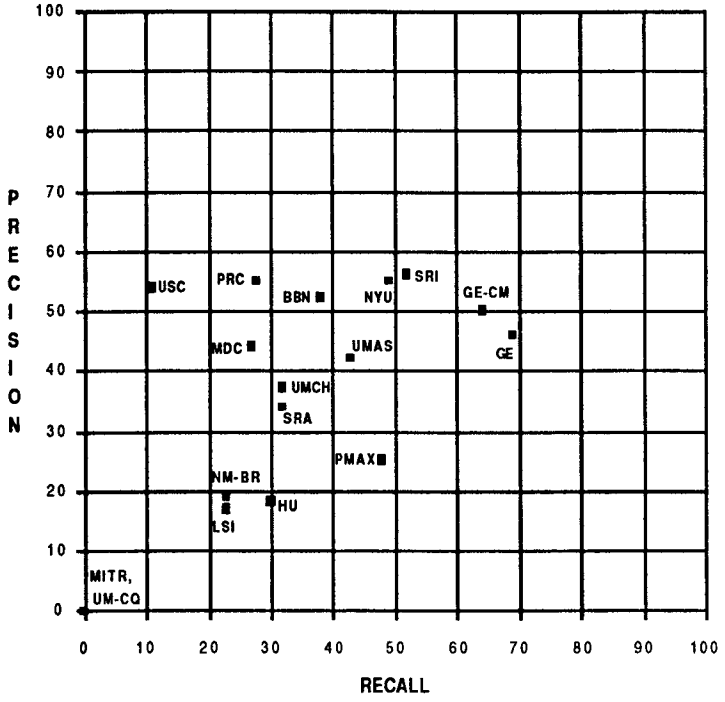

Figure H32. TST4 Base Run (Recall vs Precision): Slot 12 (phys-tgt-id)