MUC-3 EVALUATION METRICS

Nancy Chinchor, Ph.D.
Science Applications International Corporation
10260 Campus Point Drive, M/S 12
San Diego, CA 92121
(619) 458-2728

INTRODUCTION

Purpose

The MUC-3 evaluation metrics are measures of performance for the MUC-3 template fill task. Obtaining summary measures of performance necessitates the loss of information about many details of performance. The utility of summary measures for comparison of performance over time and across systems should outweigh this loss of detail. The template fill task is complex because of the varying nature of the fills for each slot and the interdependencies of the slots. The evaluation metrics used in MUC-3 were adapted from traditional measures in information retrieval and signal procesing and were still evolving to fit the more complex data extraction task of MUC-3 when the evaluation was performed. The scoring of the template fill task and the calculation of the metrics used in MUC-3 will be described here. This description is meant to assist in the analysis of the MUC-3 results and in the further evolution of the evaluation metrics.

Metrics

The measures of performance chosen for use in MUC-3 were recall, precision, fallout, and overgeneration. Recall, precision, and fallout were adapted based on their use in information retrieval. Overgeneration was developed as a measure for MUC-3. Recall is a measure of the completeness of the template fill. Precision is a measure of the accuracy of the fill. Fallout is a measure of the false alarm rate for the slots which can be filled from finite sets of slot fillers. Overgeneration is a measure of spurious generation. These measures will be described in greater detail below.

SCORE REPORT

A semi-automated scoring system was developed for MUC-3. The scoring system displayed the answer key templates, the response templates, and the messages using a flexibly customized emacs interface. During scoring, the user was asked to enter the score for displayed mismatches between the key and the response templates. Fills could generally be scored as matches, partial matches, or mismatches. Depending on the type of slot fill, the scoring system may or may not have allowed full credit to be given. The interactive scoring was carried out following well-defined scoring guidelines. Depending on the scoring guidelines, full, partial, or no credit may have been allowed for each mismatch. After the interactive scoring was complete, the scoring system produced an official score report containing template by template score reports and a summary score report for the official record. A sample summary score report produced for human comparison against the key appears in Figure 1. The following sections discuss the contents of the score report.

```
 * * * TOTAL SLOT SCORES * * *

SLOT                      POS ACT|COR PAR INC|ICR IPA|SPU MIS NON|REC PRE OVG FAL
--------------------------------+-----------+-------+-----------+----------------
template-id               118 115|114   0   0|  0   0|  1   4  39| 97  99   1
incident-date             114 110| 90  10  10| 31  10|  0   4   4| 83  86   0
incident-type             118 114|112   1   1|  0   1|  0   4   0| 95  99   0   0
category                   90 109| 88   0   0|  0   0| 21   2   7| 98  81  19  14
indiv-perps               106  61| 59   0   2| 10   0|  0  45  50| 56  97   0
org-perps                  71  68| 58   0   1| 15   0|  9  12  48| 82  85  13
perp-confidence            71  68| 56   1   2| 12   1|  9  12  48| 80  83  13   2
phys-target-ids            59  57| 54   3   0| 14   3|  0   2  77| 94  97   0
phys-target-num            41  41| 39   0   2|  0   0|  0   0  77| 95  95   0
phys-target-types          59  57| 52   4   1| 11   4|  0   2  77| 92  95   0   0
human-target-ids          145 131|129   2   0| 33   2|  2  14  23| 90  99   2
human-target-num           94  88| 79   6   2|  0   6|  1   7  23| 87  93   1
human-target-types        145 131|126   2   3| 24   2|  2  14  23| 88  97   2   0
target-nationality         35  19| 17   2   0|  3   2|  0  16 103| 51  95   0   0
instrument-types           25  22| 16   1   0|  0   0|  5   8  88| 66  75  23   0
incident-location         118 113| 88  24   1|  0   1|  0   5   0| 85  88   0
phys-effects               41  44| 37   3   0|  8   3|  4   1  89| 94  88   9   0
human-effects              56  54| 43   2   2| 10   2|  8   9  81| 78  81  15   1
--------------------------------+-----------+-------+-----------+----------------
MATCHED ONLY             1464 1402|1257  61  27|171  37| 62 119 826| 88  92   4
MATCHED/MISSING          1506 1402|1257  61  27|171  37| 62 161 857| 85  92   4
ALL TEMPLATES            1506 1420|1257  61  27|171  37| 80 161 861| 85  91   6
SET FILLS ONLY            640  618|547  16   9| 68  15| 49  68 516| 87  90   8   0
```

Figure 1: Summary Score Report

Scoring Categories

Individual slot fills in the response were scored as correct, partially correct, incorrect, noncommittal, spurious, or missing. A response was correct if it was the same as the key, partially correct if it partially approximated the key, and incorrect if it was not the same as the key. If the key and response were both blank, the response was scored as noncommittal. If the key was blank but the slot was filled, the response was scored as spurious. If the response was blank and the key was not, the response was scored as missing. Figure 2 summarizes the scoring categories relating them to the corresponding columns in the score report.

Category	Criteria	Column
Correct	response = key	COR
Partial	response ≅ key	PAR
Incorrect	response ≠ key	INC
Noncommittal	key and response are both blank	NON
Spurious	key is blank and response is not	SPU
Missing	response is blank and key is not	MIS

Figure 2: Scoring Categories

The summary score report rows show the totals for each of the categories over all templates. The slots are listed on the left hand side and the totals for each slot over all templates are given in the labeled columns. For example, the total number of physical targets correctly identified was 54. The number appears in the phys-target-ids row and the COR column of the summary score report. Note that the bottom four rows of the score report are not slot scores but rather global summary rows described in a later section.

During scoring, the scoring system automatically scored matches as correct and some partially matching hierarchically organized items as partially correct. However, many of the mismatches were interactively scored by the user. To reflect the number of items interactively scored as correct or partially correct, two columns labeled ICR and IPA were provided.

The first two columns in the score report contain the number of possible slot fills (POS) and the actual number of slot fills (ACT). The number of possible slot fills is the number of slots fills in the key plus the number of optional slot fills in the key that were matched in the response. The number of possible slot fills for each system differs depending on the optional fills given by the system. The number of actual fills given is the number of slot fillers in the response. The numbers in the possible and actual columns are used to calculate the metrics.

Calculation of Metrics

The metrics were calculated for each slot and for the summary rows. The calculations were based on information in the columns of the score report as well as on some tallies kept internally by the scoring system. The first three metrics shown in the score report are recall, precision, and overgeneration. These were calculated for each slot and were based on information contained in the score report.

Recall is a measure of completeness and was calculated as follows.

$$\text{recall} = \frac{\text{correct} + (\text{partial x } 0.5)}{\text{possible}}$$

For example, recall for the human-target-ids slot was calculated as follows.

$$\begin{aligned} \text{REC} &= \frac{\text{COR} + (\text{PAR x } 0.5)}{\text{POS}} \\ &= \frac{129 + (2 \text{ x } 0.5)}{145} \\ &= \frac{130}{145} \\ &= 0.90 \end{aligned}$$

Precision is a measure of the accuracy of the attempted fills and was calculated as follows.

$$\text{precision} = \frac{\text{correct} + (\text{ partial x } 0.5)}{\text{actual}}$$

For example, the precision for the phys-target-num slot was calculated as follows.

$$\text{PRE} = \frac{\text{COR} + (\text{ PAR X } 0.5\ )}{\text{ACT}}$$

$$= \frac{39 + (\ 0 \text{ x } 0.5\ )}{41}$$

$$= \frac{39}{41}$$

$$= 0.95$$

Overgeneration is a measure of spurious generation and was calculated as follows.

$$\text{overgeneration} = \frac{\text{spurious}}{\text{actual}}$$

For example, the amount of overgeneration in the category slot was calculated as follows.

$$\text{OVG} = \frac{\text{SPU}}{\text{ACT}}$$

$$= \frac{21}{109}$$

$$= 0.19$$

Fallout is a measure of the false alarm rate. The number of false alarms could only be measured for slots for which we knew the number of possible incorrect responses. A subset of the slots in the template fill task were filled from finite sets. The rest of the slots are filled from possibly infinite sets. Fallout measures were calculated for the finite set fill slots as follows.

$$\text{fallout} = \frac{\text{incorrect} + \text{spurious}}{\text{possible incorrect}}$$

where "possible incorrect" is the number of possible incorrect answers which could be given in the response. The number of possible incorrect is not shown in the score report but a tally is kept internally by the scoring system. The method for keeping this tally of possible incorrect has evolved during the course of the evaluation.

In order to describe this evolution, a simple calculation of fallout for a single slot in a single template will be given. The instrument type slot has 16 possible fillers. If the key contains the filler GUN and the response contains the filler GRENADE, then fallout would be

$$\begin{aligned} \text{FAL} &= \frac{\text{INC + SPU}}{\text{possible incorrect}} \\ &= \frac{1 + 0}{16 - 1} \\ &= \frac{1}{15} \\ &= 0.07 \end{aligned}$$

The number of possible incorrect is the cardinality of the set of possible answers minus the number correct in the key which is 16 - 1, or 15.

In phase one, the fallout measure assumed that the system was essentially choosing a subset of the finite set of possible fills when it gave a response. For example, if the key for the instrument type slot contained GUN and GRENADE and the response contained BOMB, GRENADE, and CUTTING DEVICE, the phase one fallout would be

$$\begin{aligned} \text{FAL} &= \frac{\text{INC + SPU}}{\text{possible incorrect}} \\ &= \frac{1 + 1}{16 - 2} \\ &= \frac{2}{14} \\ &= 0.14 \end{aligned}$$

The number of possible incorrect was the cardinality of the set minus the total number of slot fills given in the key.

During phase two, it was noticed that this simple approach to fallout was in fact erroneous for several reasons. Some finite set slots allowed multiple uses of set members due to cross-referencing requirements. For example, the slot fill CIVILIAN might be used multiple times in specifying the human target type for different human targets.

CIVILIAN: "MARIO FLORES"
CIVILIAN: "JOSE RODRIGUEZ"

Further complications arose when alternatives were given in the key for each such slot fill. In order to solve all of these problems, the calculation of the possible correct for the slot fills was revised to coincide more closely with the calculation used in information retrieval. Each separate slot fill item is now thought of as being chosen from the entire finite set of possible fill items.

In general, the number of possible incorrect is given by the following formula.

$$\sum_{keyval} (|U| - |keyval|)$$

where keyval stands for each of the key values including blanks, |U| is the cardinality of the finite set U of possible slot fillers, and |keyval| is the number of key values corresponding to the response. If there are alternative key values for a response, then |keyval| > 1. If the key is blank, then there are no corresponding key values and the contribution to the number of possible incorrect is the cardinality of the finite set.

Returning to our example of instrument types with the key containing GUN and GRENADE and the response containing BOMB, GRENADE, and CUTTING DEVICE, fallout will be recalculated using the new method of determining the possible incorrect. The number of possible incorrect is calculated by summing over the slot fills. For GUN, the number of possible incorrect is the cardinality of the set, which is 16, minus the number of slot fill alternatives given in the key, which in this case is 1. For GRENADE, the number of possible incorrect is also 15. So the number of possible incorrect for this slot is 15 + 15, or 30. Since there is 1 incorrect and 1 spurious response, fallout is 2/30, or 7%. In phase one, fallout was 14% for this same example.

If there are alternatives to a single slot fill in the key, the contribution to the number of possible incorrect by that slot fill is the cardinality of the finite set minus the number of alternatives given. For example, if the key is GUN/GRENADE, the number of possible incorrect is 16 - 2, or 14.

If the key is blank, the number of possible incorrect is the cardinality of the finite set. For example, if the instrument type slot is blank in the key and the response is GUN and GRENADE, then the fallout is

$$\begin{aligned} FAL &= \frac{INC + SPU}{\text{possible incorrect}} \\ &= \frac{0 + 2}{16} \\ &= \frac{2}{16} \\ &= 0.13 \end{aligned}$$

Notice that if the number of spurious responses is great enough, fallout can be more than 100%.

Meaning of Metrics

Recall is a measure of completeness in the sense that it measures the amount of relevant data extracted relative to the total available. It is the true positive rate. A mnemonic for recall can be constructed by imagining that you have been asked to read the entire answer key, then fill in templates with all that you have "remembered" or "recalled." Your score would be the total correctly recalled out of the total possible.

Precision is the accuracy with which a system extracts data. It is the amount of relevant data relative to the total put in by the system. A mnemonic for precision is to imagine that each time a system fills a slot it is throwing a dart at a dartboard. All of the bull's-eyes are correct. Precision is a measure of the number of bull's-eyes relative to the number of darts thrown. Precision can also be described as the tendency of a system to avoid assigning bad fillers as it assigns more good fillers.

Fallout is a measure of the false positive rate. It is the tendency of the system to assign incorrect fillers as the number of potential incorrect fillers increases. So, for a mnemonic, if you are imagining the dartboard again, fallout measures the number of darts that "fall outside" of the bull's-eye relative to the size of the area outside the bull's-eye. Fallout can only be assigned for slots with a calculable number of possible incorrect. Only some of the slots have a finite set of slot fills associated with them. The others have fills that come from potentially infinite sets and hence cannot be assigned a fallout score.

Overgeneration is a measure of spurious generation. It is the amount of spurious fillers assigned in relation to the total assigned. Overgeneration was calculated to deter overgeneration as an approach to higher scores. A mnemonic for overgeneration can be constructed by imagining that required fills and extra fills are in a box. Overgeneration is represented by the area that the extra fills take up in relation to the total area.

Summary Scores

The last four rows of the score report in Figure 1 are summary score rows. In phase one, there was one summary score row that represented the totals of the columns for the scoring categories including possible and actual. The metrics were then calculated based on those totals and appeared in the appropriate columns in the lower righthand portion of the chart. The summary metrics are always calculated from the items in the summary totals and are never the result of averaging the metrics for the slots.

In phase two, it was decided that the scoring system should keep the internal tallies needed to supply several summary score rows, only one of which would be the total of the slot scores shown in the columns of the score report. The scoring of slots in the missing and spurious templates was the issue which gave rise to multiple summary rows. In phase one, spurious templates were scored as spurious in the template id slot only. The spurious slot fillers aside from the template id slot filler were not scored as spurious. Missing templates, however, were scored in the template id slot and in the individual missing slots. This method of scoring did not penalize as much for overpopulating the database as it did for underpopulating it.

In phase two, we wanted to find out how the systems scored if overpopulating and underpopulating the database were treated equally. Two summary rows were added, one of which scored spurious and missing in the template id only and the other of which scored spurious and missing templates for all of the spurious and missing slot fills. The official scores were still taken from the same summary row as in phase one, but the other two rows were there for analysis.

The global summary rows are listed on the score report in order of strictness based on the scoring of missing and spurious templates. The MATCHED ONLY row has missing and spurious templates only scored in the template id slot. This row contains the least strict of the scores for the system. The MATCHED/MISSING row contains the official test results. The missing template slots are scored as missing. The spurious templates are scored only in the template id slot. The totals in this row are the totals of the tallies in the columns as shown. The ALL TEMPLATES row has missing template slots scored as missing and spurious template slots scored as spurious. This row contains the strictest scores for the system.

A fourth summary row was added to allow analysis of system performance on only the set fill slots. The SET FILLS ONLY row contains totals for slots with finite set fills only. A global fallout score is calculated for these slots and given in the fallout column of this row.

CONCLUSIONS AND FURTHER RESEARCH

The evaluation metrics for MUC-3 had utility for system development and for the reporting and analysis of the results of the evaluation. The metrics were adapted from simpler task models and were still evolving when the evaluation was performed. There has been consistent agreement on the necessity of basic measurements of completeness, accuracy, false alarm rate, and overgeneration. These measurements were accomplished through the metrics of recall, precision, fallout, and overgeneration as defined for MUC-3. The global summary scores provide several different views of system performance. However, further analysis of the current results is possible based on the information in the official score reports. The template by template scores are officially reported and can be used as a basis for further analysis. For example, performance at the message level can be calculated from the template by template scores for the systems.

While the metrics of recall, precision, fallout, and overgeneration have been defined for MUC-3, further research into the metrics and their implementation needs to be done. Additional measurements may be required. More refined definitions of the current measurements are probably needed. The complexities of optional fills, alternatives in the key, partial credit, and distribution of partial credit over key values, to name a few, still need to be examined more closely with consideration given to their effects on the metrics. These complexities have made it difficult to fully test the scoring system software and require more attention to be paid to detecting and isolating subtle errors. A different treatment of the slots will need to be attempted. For example, the template id slot is unique among the slots and will be kept separate when the summary measures are calculated in the future. A single overall measure of performance may be possible in the future once the roles of recall and precision are more fully determined. All of these avenues of further research have been opened up by the definition of a set of metrics for MUC-3 and the development of a scoring system embodying those metrics.

ACKNOWLEDGEMENTS

Many of the scoring issues were debated and resolved in consultation with members of the Program Committee. David Lewis provided technical guidance. Pete Halverson developed the scoring system. The participants provided feedback. Beth Sundheim was a sounding board for many of the issues that arose during the evolution of the MUC-3 scoring.