MUC-7 TEST SCORES

This appendix contains the summary score reports for each system in the MUC-7 evaluation. The reports are ordered by task (Template Element, Template Relation, Scenario Template, Named Entity, and Coreference) and secondarily by site and language (in alphabetical order).

A brief introduction to reading the MUC-7 score reports is presented here. First, the common aspects of the score reports will be presented followed by the task-specific aspects. Then references will be given for more detailed information on the metrics, scoring algorithms, and statistical significance testing.

The first column in the report, SLOT, contains one of the following types of label:

A slot name, e.g. **ent_name**

An object name, e.g. **entity**

The following tallies for the test set are presented in the lefthand columns of the score report:

POS Number possible (**COR + INC + MIS**), i.e., the number of fills in the answer key plus any optional fills allowed by the key and generated by the system

ACT Number actual (**COR + INC + SPU**),i.e., the number of fills generated by the system under evaluation

COR Number correct

PAR Number partially correct (no partial credit was given in MUC-7)

INC Number incorrect

MIS Number missing

SPU Number spurious

NON Number non-committal (null fills generated by system that were also null in the answer key)

The righthand columns in the report contain the values for the following extraction metrics:

REC Recall **= COR/POS**

PRE Precision **= COR/ACT**

UND Undergeneration **= MIS/POS**

OVG Overgeneration **= SPU/ACT**

SUB Substitution **= INC/(COR + INC)**

ERR Error per response fill **= (INC + SPU + MIS)/(COR + INC + SPU + MIS)**

The first row at the bottom of the report is marked as the **ALL SLOTS** row and gives the official overall tallies and the overall **REC**, **PRE**, **UND**, **OVG**, **SUB**, and **ERR**. The tallies in the **POS**, **ACT**, **COR**, **PAR**, **INC**, **SPU**, **MIS**, and **NON** columns add up to the **ALL SLOTS** values, but not all rows are considered. The rows that are not considered are the **OBJ SCORES** (i.e., the section where the entries in the **SLOT** column are object names). The scores for the metrics in the **ALL SLOTS** row are not computed by averaging the scores in the individual slot rows, but instead are computed on the basis of the tally totals contained in the **ALL SLOTS** row itself.

The remaining row at the bottom of the report contains scores for the following metrics:

F-MEASURES F-Measures (weighted combination of recall and precision scores in **ALL SLOTS** row)

F-Measure scores corresponding to three different weightings of precision relative to recall are computed. The **P&R** value weights precision the same as recall; the **2P&R** value gives twice the weight to precision; the **P&2R**

value weights precision half as much as recall. The general formula is:

$$F = \frac{(\beta^2 + 1.0) \times P \times R}{(\beta^2 \times P) + R}$$

Information Extraction

Scenario Template

In the Scenario Template task, there is one template object per article and the content slot may or may not be filled depending on whether the text contains a relevant event. The ability of the system to judge relevancy is reflected in the TEXT-FILTERING row of the score report.

TEXT-FILTERING	Text filtering (recall and precision for document detection)

The score category columns used to contain the tallies for these decisions across the test set are as follows: **COR** (decides relevant and relevant is correct), **SPU** (decides relevant and nonrelevant is correct), **MIS** (decides nonrelevant and relevant is correct), and **NON** (decides nonrelevant and nonrelevant is correct). Text filtering formulas are therefore as follows: **REC = COR/(COR + MIS); PRE = COR/(COR + SPU); UND = MIS/(COR + MIS); OVG = SPU/(COR + SPU)**.

Template Element and Template Relation

The variation on the Scenario Template (**ST**) score report for Template Element (**TE**) and Template Relation(**TR**) is minor; there is no template object and no text filtering score in either because the objects are generated independent of any domain (scenario) relevance criteria.

Named Entity

For the Named Entity (**NE**) task, the reporting is slightly more extensive. The first table is the same as for Information Extraction, i.e., it contains subtask scores. The second table reports the scores according to which part of the document the named entity appeared in. The remaining tables are the same as those in the Information Extraction tasks.

Coreference

The Coreference (**CO**) task is quite different from the other four tasks and the output is also different. The number of coreference chains per document is given for the key and the response. The scoring method looks at linkages that define equivalence classes, and the important measure is how many linkages need to be added or taken away to get the correct equivalence class. The score reports contain recall, precision, and F-measure scores. Please note that the recall and precision scores for **CO** are reported as both ratios and percentages.

REFERENCES

- Chinchor, N.; Hirschman, L.; and Lewis, D. D. (1993). “Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3).” *Computational Linguistics*, 19(3),409 - 449.
- Chinchor, N. (1992). “The Statistical Significance of the MUC-4 Results.” In *Proceedings, Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufmann. San Mateo, CA.
- Vilain, M.; Burger, J.; Aberdeen, J.; Connolly, D.; and Hirschman, L. (1995). “A Model-Theoretic Coreference Scoring Scheme.” In *Proceedings, Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufmann. San Mateo, CA.