# USC:
# MUC-4 Test Results and Analysis

*D. Moldovan, S. Cha, M. Chung, K. Hendrickson, J. Kim, and S. Kowalski*

Parallel Knowledge Processing Laboratory
University of Southern California
Los Angeles, California 90089-2562
moldovan@gringo.usc.edu
(213)740-4477

## INTRODUCTION

The University of Southern California is participating, for the first time, in the message understanding conferences. A team consisting of one faculty and five doctoral students started the work for MUC-4 in January 1992. This work is an extension of a project to build a massively parallel computer for natural language processing called Semantic Network Array Processor (SNAP).

## RESULTS

### Scoring Results

During the final week of testing, our system was run on test sets TST3 and TST4. Test set TST3 contains 100 articles from the same time period as the training corpus (DEV), and test sets TST1 and TST2. The summary of score results for TST3 is shown in Table 1. Test set TST4 contains 100 articles from a different time period then those of TST3. The summary of score results for TST4 is shown in Table 2. The complete score results for TST3 and TST4 can be found in Appendix G.

### Recall

The recall metric (REC column in Tables 1 and 2) is a measure of the system's ability to extract relevant information from the text. For the TST3 test set, our recall score was 7% as shown in the ALL TEMPLATES and MATCHED/MISSING rows of Table 1. If missing templates are disregarded, our recall score for TST3 improves to 30% as is shown in the MATCHED/SPURIOUS and MATCHED ONLY rows of Table 1. For the TST4 test set, our recall score was 12% as shown in the ALL TEMPLATES and MATCHED/MISSING rows of Table 2. If missing templates are disregarded, our recall score for TST4 improves to 31% as is shown in the MATCHED/SPURIOUS and MATCHED ONLY rows of Table 2.

### Precision

| SLOT | POS | ACT | COR | PAR | INC | ICR | IPA | SPU | MIS | NON | REC | PRE | OVG | FAL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MATCHED/MISSING | 1508 | 174 | 85 | 31 | 26 | 4 | 11 | 32 | 1366 | 1142 | 7 | 58 | 18 | |
| MATCHED/SPURIOUS | 332 | 637 | 85 | 31 | 26 | 4 | 11 | 495 | 190 | 1110 | 30 | 16 | 78 | |
| MATCHED ONLY | 332 | 174 | 85 | 31 | 26 | 4 | 11 | 32 | 190 | 148 | 30 | 58 | 18 | |
| ALL TEMPLATES | 1508 | 637 | 85 | 31 | 26 | 4 | 11 | 495 | 1366 | 2104 | 7 | 16 | 78 | |
| SET FILLS ONLY | 719 | 89 | 46 | 16 | 14 | 0 | 1 | 13 | 643 | 537 | 8 | 61 | 15 | 0 |
| STRING FILLS ONLY | 390 | 48 | 20 | 5 | 7 | 1 | 5 | 16 | 358 | 320 | 6 | 47 | 33 | |
| | | | | | | | | | P&R | | 2P&R | | | P&2R |
| F-MEASURES | | | | | | | | | 9.74 | | 12.73 | | | 7.89 |

**Table 1:** Summary of Score Results for TST3.

| SLOT | POS | ACT | COR | PAR | INC | ICR | IPA | SPU | MIS | NON | REC | PRE | OVG | FAL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MATCHED/MISSING | 1105 | 208 | 124 | 40 | 30 | 8 | 23 | 14 | 911 | 745 | 13 | 69 | 7 | |
| MATCHED/SPURIOUS | 456 | 508 | 124 | 40 | 30 | 8 | 23 | 314 | 262 | 844 | 32 | 28 | 62 | |
| MATCHED ONLY | 456 | 208 | 124 | 40 | 30 | 8 | 23 | 14 | 262 | 236 | 32 | 69 | 7 | |
| ALL TEMPLATES | 1105 | 508 | 124 | 40 | 30 | 8 | 23 | 314 | 911 | 1353 | 13 | 28 | 62 | |
| SET FILLS ONLY | 538 | 115 | 78 | 20 | 10 | 0 | 6 | 7 | 430 | 339 | 16 | 76 | 6 | 0 |
| STRING FILLS ONLY | 288 | 50 | 25 | 6 | 13 | 0 | 6 | 6 | 244 | 209 | 10 | 56 | 12 | |
| | | | | | | | | | P&R | | 2P&R | | | P&2R |
| F-MEASURES | | | | | | | | | 17.76 | | 22.75 | | | 14.56 |

**Table 2:** Summary of Score Results for TST4.

The precision metric (PRE column in Tables 1 and 2) is a measure of the correctness of the system's output. For the TST3 test set, our precision score was 16% as shown in the ALL TEMPLATES and MATCHED/SPURIOUS rows of Table 1. If spurious templates are disregarded, our precision score for TST3 improves to 58% as is shown in the MATCHED/MISSING and MATCHED ONLY rows of Table 1. For the TST4 test set, our precision score was 26% as shown in the ALL TEMPLATES and MATCHED/SPURIOUS rows of Table 2. If missing templates are disregarded, our precision score for TST4 improves to 69% as is shown in the MATCHED/MISSING and MATCHED ONLY rows of Table 2.

### Analysis of Results

The large disparity of scores between TST3 and TST4 can be partially attributed to the ability of our system to generate the required templates with enough correct slots that they can exceed the minimum matching criteria of the scoring software. For TST3, we only generated 16 templates out of the 103 possible. 61 of our templates were spurious. We did much better with TST4, in that we generated 24 of the 71 possible templates and had only 41 spurious templates.

## LEVEL OF EFFORT

The total effort for MUC-4 is estimated at approximately 1,450 hours. This breaks down as follows:

| | |
|---|---|
| Knowledge base construction | 25% |
| Preprocessor | 15% |
| Memory based parser | 25% |
| Template generation | 20% |
| System integration | 10% |
| Scoring procedure | 5% |

## LIMITING FACTORS

The main limiting factor for us was that we started almost from scratch. We did not have a lexicon, parser, knowledge base, nor an inference engine; we only had ideas and a small parser which turned out to be useless for this large application. As our knowledge base grew we started to run out of memory in the parallel computer's controller board, so we had to redesign this board. Since it was not ready in time to be useful for MUC-4 testing, we ended up using the software simulator of the parallel computer which was very slow. It takes more than one hour to process a message using the simulator, but only seconds when using the actual parallel computer.

Regarding the limiting factors in performance of the system we have noticed that: (1) our discourse processing capability was insufficient, (2) the lexicon was too small, (3) the parser does not address enough linguistic problems, (4) more basic concept sequences are needed, and (5) more inferencing rules are needed.

Although the MUC-4 experiment presented many challenging problems, we have not yet reached the limit of our technology. We built the system using only one test message, and only had a working system starting in April. The last month was used to fine tune the system using all 100 messages in the previous corpus.

## STRENGTHS AND WEAKNESSES

### Strengths

Memory based parsing seems powerful and offers many advantages. The use of integrated semantic and syntactic parsing was successful. The structure of the knowledge base and the dynamic combination of various concept sequences to handle arbitrary input sentences worked well.

### Weaknesses

Because of insufficient concept sequences in the knowledge base, the parser's output is mostly a syntactic description of the sentences, as opposed to a semantic description. The template generator doesn't yet do any discourse processing. High-level inferencing is needed. The knowledge base was built to work with the parser, without much regard for the inferencing process.

## REUSABILITY

Assuming that the domain and the required output is changed, approximately 75% of the knowledge base and the lexicon is reusable. None of the inferencing rules for filling templates are reusable, although some of the structure might be reusable.

## WHAT WAS LEARNED

We have come to a greater appreciation of how complex the problem really is. Further improvements of the system need to focus on discourse processing and high-level inferencing. Also, common-sense knowledge must be added to the knowledge base, and parallel inferencing methods must be developed to apply this knowledge. We also see a great need for automating the construction and enhancement of the knowledge base.

Over all, our experience with MUC-4 has been useful and rewarding. More than anything, it has focused our work.

## ACKNOWLEDGEMENT

We are grateful to Richard Tong from ADS for making available to us part of the dictionary and taxonomy, and to Beth Sundheim for facilitating this. This work was partially funded by the National Science Foundation under grant #MIP-9009109.

# PART III: SYSTEM DESCRIPTIONS

The papers in this section, which were prepared by each of the sites that completed the MUC-4 evaluation, describe the systems that were tested. The papers are intended not only to outline each system's architecture but also to provide the reader with an understanding of the effectiveness of the techniques that were used to handle the particular phenomena found in the MUC-4 corpus. To make the discussion of these techniques concrete, most of the sites make specific reference to some of the phenomena found in message TST2-MUC4-0048 from the dry-run test set and discuss their system's handling of those phenomena. The full text and answer key templates for that message are found in appendix F of the proceedings.

The sites were asked to include the following pieces of information in this paper:

* Background: how/for what the system was developed, and how much time was spent on the system *before* MUC-4

* Explanation of the modules of the system

* Explanation of flow of control (interleaved/sequential/...)

* Explanation (without system-specific jargon) of processing stages:
  - Identification of relevant texts and paragraphs
  - Lexical look-up (example of output and lexicon)
  - Syntactic analysis (example of output and grammar)
  - Semantic analysis (example of output and semantic rules)
  - Reference resolution
  - Template fill

* Sample filled-in template, with an explanation of interesting things:
  - things system got right
  - things system got wrong