# SRI INTERNATIONAL'S TACITUS SYSTEM: MUC-3 TEST RESULTS AND ANALYSIS

*Jerry R. Hobbs*
SRI International
Menlo Park, California 94025
hobbs@ai.sri.com
(415) 859-2229

## RESULTS

This site report is intended as a companion piece to the System Summary appearing in this volume and is best read in conjunction with it. In particular, it refers to the various modules of the system which are described in that paper.

Here only the overall results will be summarized. A more detailed, component-by-component analysis of the results is contained in the System Summary.

Our results for the TST2 corpus were as follows:

| | Recall | Precision |
|---|---|---|
| Matched Templates | 44 | 65 |
| Matched/Missing | 25 | 65 |
| All Templates | 25 | 48 |

Our precision was the highest of any of the sites. Our recall was somewhere in the middle. It is as yet unclear whether high recall-high precision systems will evolve more rapidly from low recall-high precision systems or high recall-low precision systems.

The significant drop in recall we experienced from Matched Templates Only to Matched/Missing is an indication that we were failing on messages with a large number of template entries. Much of this is probably due to failures in handling lists of names, and could be improved by specialized handling of this phenomenon.

We also ran our system, configured identically to the TST2 run, on the first 100 messages of the development set. The results were as follows:

| | Recall | Precision |
|---|---|---|
| Matched Templates | 46 | 64 |
| Matched/Missing | 37 | 64 |
| All Templates | 37 | 53 |

Here recall was considerably better, as would be expected since the messages were used for development.

While there are a number of parameter settings possible in our system, we decided upon optimal values, and those values were used. An explanation of the parameters and how we decided what was optimal is too detailed and system-particular for this report. None of the decisions was made on the basis of total recall and precision on a test set. All the decisions were made on a much more local basis.

## LEVEL OF EFFORT

The only way of even approximating the amount of time spent on this effort is from figures on time charged to the project. All participants in the MUC-3 process will realize that this is not a very reliable way of estimating the level of effort.

Since the preliminary MUC-3 workshop in February, approximately 800 person-hours were spent on the project.

The only possible way to break that down into subtasks is by personnel.

| | |
|---|---|
| Preprocessor, system development, testing: | 180 hours |
| Development of parsing algorithms: | 180 hours |
| Grammar development: | 220 hours |
| Pragmatics and template-generation: | 220 hours |

## THE LIMITING FACTOR

Time.

## TRAINING

The amount of the training corpus that was used varied with the component. For the relevance filter, all 1400 available messages were used. For the lexicon, every word in the first 600 and last 200 messages and in the TST1 corpus were entered. For the remaining messages, those words occurring more than once and all non-nouns were entered.

For syntax and pragmatics, we were able only to focus on the first 100 messages in the development corpus.

Tests were run almost entirely on the first 100 messages because those were the only ones for which a reliable key existed and because concentrating on those would give us a stable measure of progress.

The system improved over time. On the February TST1 run, our recall was 14% and our precision was 68% on Matched and Missing Templates. At the end of March, on the first 100 messages in the development set, our recall was 22% and our precision was 63%. At the time of the TST2 evaluation, on the first 100 messages in the development set, our recall was 37% and our precision was 64%.

## WHAT WAS AND WAS NOT SUCCESSFUL

As described in the System Summary, we felt that the treatment of unknown words was for the most part adequate.

The statistical relevance filter was extremely successful. The keyword antifilter, on the other hand, is apparently far too coarse and needs to be refined or eliminated.

We felt syntactic analysis was a stunning success. At the beginning of this effort, we despaired of being able to handle sentences of the length and complexity of those in the MUC-3 corpus, and indeed many sites abandoned syntactic analysis altogether. Now, however, we feel that the syntactic analysis of material such as this is very nearly a solved problem. The coverage of our grammar, our scheduling parser, and our heuristic of using the best sequence of fragments for failed parses combined to enable us to get a very high proportion of the propositional content out of every sentence. The mistakes that we found in the first 20 messages of TST2 can, for the most part, be attributed to about five or six causes, which could be remedied with a few days work.

On the other hand, the results for terminal substring parsing, our method for dealing with sentences of more than 60 words, are inconclusive, and we believe this technique could be improved.

In pragmatics, much work remains to be done. A large number of fairly simple axioms need to be written, as well as some more complex axioms. In the course of our preparation for MUC-2 and MUC-3, we have made sacrifices in robustness for the sake of efficiency, and we would like to re-examine the trade-offs. We would like to push more of the problems of syntactic and lexical ambiguity into the pragmatics component, rather than relying on syntactic heuristics. We would also like to further constrain factoring, which now sometimes results in the incorrect identification of distinct events.

In template-generation, we feel our basic framework is adequate, but a great many details must be added.

The module we would most like to rewrite is in fact not now a module but should be. It consists of the various treatments of subcategorization, selectional constraints, generation of canonical predicate-argument relations, and the sort hierarchy in pragmatics. At the present time, due to various historical accidents and compromises, these are all effectively separate. The new module would give a unified treatment to this whole set of phenomena.

## USABILITY FOR OTHER APPLICATIONS

In the preprocessor, the spelling corrector and the morphological word assignment component are usable in other applications without change.

The methods used in the relevance filter are usable in other applications, but, of course, the particular statistical model and set of keywords are not.

In the syntactic analysis component, the grammar and parsing programs and the vast majority of the core lexicon are usable without change in another application. Only about five or six grammar rules are particular to this domain, encoding the structure of the heading, interview conventions, "[words indistinct]", and so on. The logical form produced is application-independent.

The theorem prover on which the pragmatics component is based is application-independent. All of the enhancements we have made in our MUC-3 effort would have benefited our MUC-2 effort as well.

In the knowledge base, only about 20 core axioms carried over from the opreps domain to the terrorist domain. Since most of the current set of axioms is geared toward MUC-3's particular task, there would very probably not be much more of a carry-over to a new domain.

The extent to which the template-generation component would carry over to a new application depends on the extent to which the same baroque requirements are imposed on the output.

## WHAT WAS LEARNED ABOUT EVALUATION

On the one hand, the mapping from texts to templates is discontinuous in the extreme. One mishandled semicolon can cost 4% in recall in the overall score, for example. Therefore, the numerical results of this evaluation must be taken with a grain of salt. Things can be learned about the various systems only by a deeper analysis of their performance. On the other hand, the task is difficult enough to provide a real challenge, so that pushing recall and precision both into the 70s or 80s will require the system to do virtually everything right.

Leading up to MUC-3 there were a great many difficulties to be worked out, diverting the attention of researchers from research to the mechanics of evaluation. It is to be hoped that most of these problems have been settled and that for MUC-4 they will constitute less of a drain on researchers' time.

We feel the task of the MUC-3 evaluation is both feasible and challenging in the relatively short term. How practical it is is for others to judge.

## ACKNOWLEDGEMENTS

This research has been funded by the Defense Advanced Research Projects Agency under Office of Naval Research contracts N00014-85-C-0013 and N00014-90-C-0220.