THE PRC PAKTUS SYSTEM:
MUC-4 TEST RESULTS AND ANALYSIS

Bruce Loatman

PRC Inc.
Technology Division
1500 PRC Drive, McLean, VA 22102
loatman_bruce@po.gis.prc.com

INTRODUCTION

In commenting on PRC's performance in MUC-3, we reported [1] that the component of our system that could most improve performance was discourse analysis. The MUC-4 exercise has strongly confirmed that view. We added a new discourse module to PAKTUS, made very few changes to the other system components, and the result was significantly improved performance. This paper discusses our test results, what we focused on for this task, what worked well, and what would improve performance further. A companion paper in this volume describes how our system functions.

KEY SYSTEM FEATURES

The PRC PAKTUS system used for MUC-4 is essentially the same linguistic system that we used for MUC-3, with the addition of a generic discourse analysis module. PAKTUS applies lexical, syntactic, semantic, and discourse analysis to all text in each document. The linguistic modules for this are nearly independent of the task domain (i.e., MUC-4, but some of the data – lexical entries, and a few grammar rules – are tuned for the MUC-4 text corpus). Task-specific template filling and filtering operations are performed only after linguistic analysis is completed.

The task-specific patterns that determine what to extract from the discourse structures were only minimally defined due to the limited time and effort available. The other task-specific additions to the system were the location set list, and functions for better recognizing time and location of events.

RESULTS

Figure 1 summarizes PRC's scores for MUC-4. The scoring notation is explained in Appendix G. Overall, we were pleased with the performance improvement since MUC-3, which was obtained with only about 4 person months of linguistic development effort, little of which was specific to the MUC-4 task. The most significant new development, compared to our MUC-3 system, is the addition of the discourse analysis module. This module is generic for expository discourse such as is found in news reports. Application-specific extraction requirements are maintained separately from the discourse module, are applied only after it executes, and were minimally specified for MUC-4.

Our system generally had much better precision than recall in these tests. We expected this because it uses complete linguistic analysis designed for text understanding, and because it has only a very limited amount of task-specific knowledge. For example, its discourse analysis module was trained on only 8 of the MUC-4 pre-test corpus of 1500 reports. For these same reasons, we also expected a high degree of corpus independence, and this was supported by the similarity of scores on TST3 and TST4.

The main limiting factors for PRC were time and availability of people for development. We directed most of our energies to generic linguistic development, and the linguistic aspects of the task have essentially been completed. Because we had little time remaining to devote to MUC-4-specific issues, however, much of the information that PAKTUS produced through syntactic, semantic, and discourse analysis did not find its way into the template fills.

```
                       POS  ACT|COR PAR  INC|ICR  IPA|SPU MIS  NON|REC  PRE  OVG  FAL
-------------------------------+------------+--------+------------+---------------
TST3
MATCHED/MISSING       1552  641|364 128   85|  8   73| 64 975 1166| 28   67   10
MATCHED/SPURIOUS      1009 1042|364 128   85|  8   73|465 432 1217| 42   41   45
MATCHED ONLY          1009  641|364 128   85|  8   73| 64 432  625| 42   67   10
ALL TEMPLATES         1552 1042|364 128   85|  8   73|465 975 1758| 28   41   45
SET FILLS ONLY         742  303|199  39   39|  0   20| 26 465  546| 29   72    8    0
STRING FILLS ONLY      401  129| 62  19   24|  7   19| 24 296  328| 18   55   19
TEXT FILTERING          66   56| 48   *    *|  *    *|  8  18   26| 73   86   14   24
F-MEASURES                           P&R        2P&R     P&2R
All Templates                      33.28        37.52    29.90
1ST                                27.87        24.02    33.17
1MT                                39.39        47.02    33.89
NST                                49.52        57.78    43.33
2MT                                26.26        27.84    24.86
-------------------------------+------------+--------+------------+---------------
TST4
MATCHED/MISSING       1155  480|279  83   78| 29   38| 40 715  753| 28   67    8
MATCHED/SPURIOUS       703  803|279  83   78| 29   38|363 263  889| 46   40   45
MATCHED ONLY           703  480|279  83   78| 29   38| 40 263  403| 46   67    8
ALL TEMPLATES         1155  803|279  83   78| 29   38|363 715 1239| 28   40   45
SET FILLS ONLY         566  239|145  29   44|  7   14| 21 348  337| 28   67    9    0
STRING FILLS ONLY      298   98| 62   9   15| 13    9| 12 212  214| 22   68   12
TEXT FILTERING          54   51| 39   *    *|  *    *| 12  15   34| 72   76   24   26
F-MEASURES                           P&R        2P&R     P&2R
All Templates                      32.94        36.84    29.79
```

Figure 1. PRC Score Summary

DEVELOPMENT EFFORT

Three PRC researchers participated in linguistic development that contributed to MUC-4 performance. Most of this development was generic, however, and will support applications other than MUC-4. Figure 2 shows an estimate of our level of effort broken down by linguistic task. Our total linguistic development effort was about four months, with almost 40% of that on discourse analysis. Significant effort also went into time and location grammar functions, although this is small compared to the prior effort that went into the overall grammar.

Lexicon entry was minimal, consisting primarily of semi-automatic entry of the MUC-4 location set list. Many words from the MUC-4 corpus have never been entered into the PAKTUS lexicon. Instead, heuristics based on word morphology make guesses about these unrecognized words.

The specific changes and additions to the PAKTUS knowledge bases for MUC-4 are enumerated in Figure 3. Most of the lexical additions were from the MUC-4 location set list. These were added semi-automatically in batch mode. Other lexical additions were based on short lists of exceptions to our unknown word heuristics, derived by scanning traces from the entire 1500 document MUC-4 pre-test corpus.

– Discourse	1.65 months
– Output Template and Format	0.25
– Lexicon Entry	0.25
– Time & Location grammar	1.50
– Preprocessor	0.16
– Lexicon Problem Identification	0.25
– Other Troubleshooting	0.31
TOTAL	4.37 months

Figure 2. Breakdown of Linguistic Development Efforts

One notable area that would have significantly improved performance was the definition of MUC-4-specific conceptual patterns. These are used to extract information from the discourse structures. Very little was done here, however, due to limited time and resources. Only 88 of these patterns were added. We had intended to define several hundred, but that would have required about another month of effort.

Knowledge Type	Core System	New/ Mod for MUC-4
Words (Stems)	6,314	2,436
Tokens	10,366	2,532
Compounds	237	360
Idioms	71	2
Verb categories	16	0
Nominal categories	407	0
Adverb categories	10	0
Closed categories	41	0
Grammar Arcs	265	4
Grammar States	78	0
Concepts	386	0
Subconcepts	18	0
Conceptual Patterns	31	88
Domain Template	0	1

Figure 3. Additions/ Modifications to PAKTUS Knowledge Bases for MUC-4

SYSTEM TRAINING AND PERFORMANCE IMPROVEMENT

As already noted, the most significant system improvement was in discourse analysis. The new discourse module was trained on only 8 documents from the test2 set. These were documents 1, 3, 10, 11, 48, 63, 99, and 100. The time and location grammar and functional changes were based on manual analysis of the 100 test2 documents. The entire pre-test corpus was scanned automatically to identify words missing from our lexicon, but only a few of these were entered – those more common words that did not conform to our unrecognized word heuristics.

The improvement in PAKTUS's linguistic performance from MUC-3 up to the day of testing for MUC-4 can be seen in Figure 4, derived from the test runs on the test2 corpus, using the F-measure specified for MUC-4. The development was carried out during April and May, 1992.

The basic functionality of the new discourse module was completed on May 6, and it dramatically improved performance. This module has two main functions: 1) it builds discourse topic structures, and 2) it unifies noun phrases that refer to the same entity. There is a rather intricate interaction between these two functions, and this had to be carefully developed over the next ten days (through May 17), so that improvement in one function did not impair the other.

After completion of the two basic discourse functions, enhancements (pronoun reference, etc.) were added to the discourse module, through May 25. This allowed only three days for MUC-4-specific knowledge to be added that could take advantage of the new discourse module.

It can be seen from figure 4 that, once the discourse functions were properly integrated (on May 17), performance improvement averaged one point per day over the last eleven days before official MUC-4 testing. We believe that the system is far from the limit of its extraction capability based on its existing linguistic components. This belief is supported by the ease with which we improved performance on the MUC-4 conference walkthrough document (test2, document 48) by adding a few MUC-4-specific conceptual patterns.

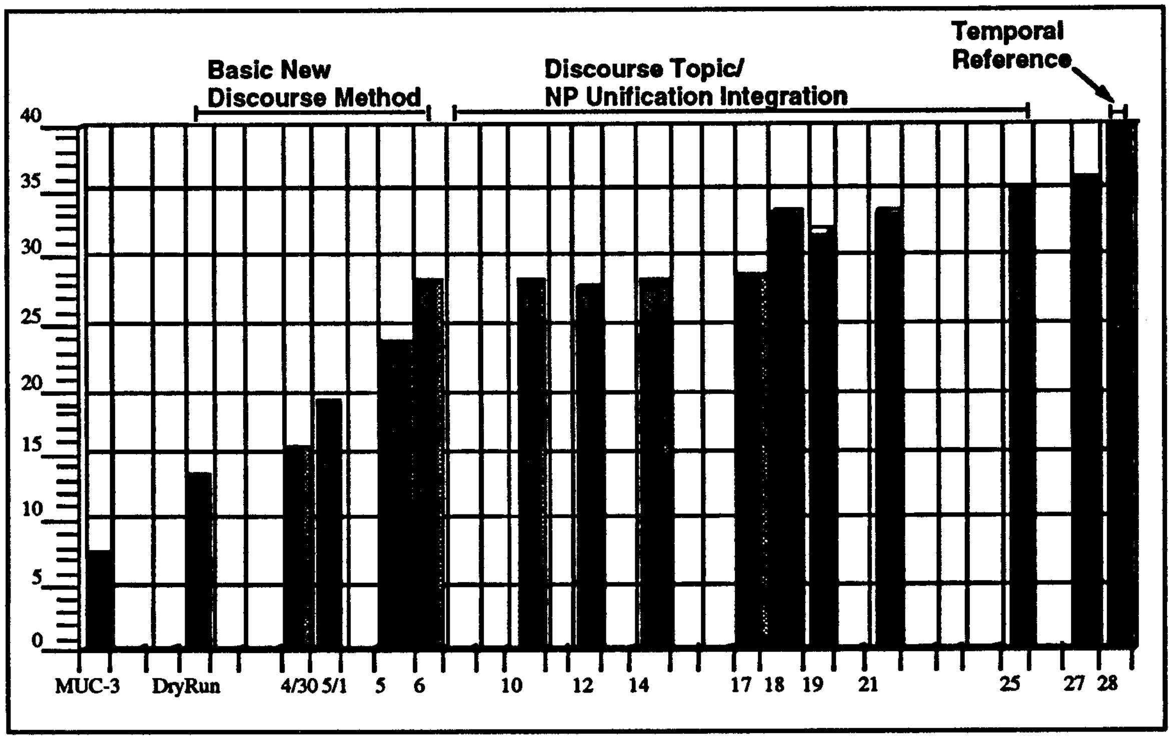

Figure 4. Performance Improvement During Development

REUSABILITY OF THE SYSTEM

Almost all of PAKTUS is generic and can be applied to other applications. All of its processes, including the new discourse analysis module, are generic. They operate on a set of object-oriented knowledge bases, some of which are generic (common English grammar and lexicon) and some of which are domain-specific (conceptual templates).

The primary tasks in applying PAKTUS to a new domain or improving its performance in an existing domain, are lexicon addition and conceptual template specification, both of which are relatively easy (compared to changing the grammar, for example).

Two other tasks that must be done, but only once for each new domain, are specifying the input document formats, and the output specifications. These are template-driven in PAKTUS.

For MUC-4 we used the template supplied by NRaD, adding a function for each template slot to gather information from our generic discourse data structures.

WHAT WE LEARNED

About PAKTUS

We learned that the current implementation of PAKTUS, including the new discourse module, is robust and adaptable. The more complex components (syntactic, semantic, and discourse analysis modules) are stable and competent enough to apply the system to different domains and produce useful results, by adding domain-specific knowledge (lexicon and conceptual patterns). We were particularly pleased to learn that it was not necessary to manually analyze much of the corpus in detail. This was done for only eight documents for MUC-4. The full development corpus was used only for lexicon development and testing the system for overall performance and logic errors.

About the Task

MUC-4 reinforced our appreciation of the importance of clearly defined output specifications, and the utility of having answer keys against which to measure the system's progress. We are already using the MUC-4 task specifications as a model for a new application of our system.

We have also come to appreciate the utility of an automated scoring program to the development effort. This quickly eliminates much uncertainty about whether a new development is useful or not, and thereby speeds system development.

About Evaluation

It is difficult to define evaluation measures for a task of this nature. Although the MUC-4 measures seem better than those of MUC-3, they do not accurately convey the true performance in some situations. For example, the system might correctly fill in 75% of the information for a template, but not report it because it got the wrong date (events over three months old are not reported), or the wrong country. We would prefer to report all incidents, with an extra slot indicating whether they are considered relevant or not. This seems more appropriate for evaluating linguistic competence. We also suspect that many analysts using such a system would like to be able to identify "irrelevant" incidents, especially since, given the current limits of linguistic technology, they may be relevant after all.

REFERENCE

[1] Kariya, C, "PRC PAKTUS: MUC-3 Test Results and Analysis", *Proceedings of the 3rd Message Understanding Conference*, San Mateo, CA: Morgan Kaufmann, 1991.