UNISYS:
MUC-3 TEST RESULTS AND ANALYSIS

Carl Weir, Robin McEntire, Barry Silk, and Tim Finin
Unisys Center for Advanced Information Technology
Paoli, Pennsylvania
weir@prc.unisys.com
(215) 648-2369

INTRODUCTION

The Unisys MUC-3 system is based on a three-tiered approach to text processing in which a novel and quite powerful knowledge-based form of information retrieval plays a central role. The main components of this approach are as follows:

A Keyword-Based Information Retrieval Component.
This component predicts the occurrence of types of events in texts based on the presence of key words and phrases.

A Knowledge-Based Information Retrieval Component.
This component, called KBIRD in the Unisys MUC-3 system, performs the following tasks:

- Based on the co-occurrence of the predictions made by the keyword-based analysis component and expressions and concepts discovered in a given text, it predicts the likely occurrence of additional event types.
- It locates instances of predicted event types in texts.
- It identifies possible slot values for located instances of events.

[A Linguistic Analysis Component.]
Although a natural language processing component was included in the design of the Unisys MUC-3 system as a third level of text analysis, not enough time was available during the MUC-3 development cycle both to develop a knowledge-based information retrieval component and to port the Unisys Pundit text-processing system to the MUC-3 terrorist domain. A decision was made to focus on developing the knowledge-based information retrieval component and postpone the integration of Pundit until MUC-4.

A Template Generation Component.
An application-specific Prolog program was written to merge templates describing the same event, and to select the most likely slot values for templates in cases where multiple slot values were proposed.

The Unisys MUC-3 development effort was comprised of two full-time Unisys staff members and one government employee on industrial rotation. A total of 2650 person-hours were put into the project, 800 of which were contributed by the government employee. The effort was partially supported by a DARPA grant, which covered approximately 30% of the development cost.[1] The bulk of the effort involved the development of the KBIRD system and its MUC-3 rule base. These two tasks took approximately the same amount of time, and in total comprised roughly 85% of the effort.

[1] Work on this project was partially supported by Darpa under contract MDA-903-89-C-0041.

SLOT	POS	ACT	COR	PAR	INC	ICR	IPA	SPU	MIS	NON	REC	PRE	OVG	FAL
template-id	104	47	38	0	0	0	0	9	66	36	36	81	19	
incident-date	100	35	25	7	3	0	7	0	65	4	28	81	0	
incident-type	104	38	34	4	0	0	4	0	66	0	35	95	0	0
category	71	28	15	0	7	0	0	6	49	27	21	54	21	10
indiv-perps	93	14	6	1	3	0	1	4	83	41	7	46	28	
org-perps	60	41	11	1	6	2	1	23	42	37	19	28	56	
perp-confidence	60	37	7	2	9	1	2	19	42	37	13	22	51	5
phys-target-ids	53	17	10	2	0	0	2	5	41	66	21	65	29	
phys-target-num	37	11	7	0	4	0	0	0	26	67	19	64	0	
phys-target-types	53	17	9	1	2	0	1	5	41	66	18	56	29	0
human-target-ids	127	37	12	7	6	3	7	12	102	19	12	42	32	
human-target-num	83	25	10	1	12	0	1	2	60	19	13	42	8	
human-target-types	127	37	15	7	3	3	7	12	102	19	14	50	32	1
target-nationality	16	3	0	0	0	0	0	3	16	89	0	0	100	0
instrument-types	23	13	3	1	0	0	0	9	19	77	15	27	69	0
incident-location	104	33	8	22	3	0	2	0	71	0	18	58	0	
phys-effects	36	17	6	2	1	1	2	8	27	77	19	41	47	1
human-effects	53	1	0	0	1	0	0	0	52	71	0	0	0	0
MATCHED ONLY	519	451	216	58	60	10	37	117	185	274	47	54	26	
MATCHED/MISSING	1304	451	216	58	60	10	37	117	970	752	19	54	26	
ALL TEMPLATES	1304	558	216	58	60	10	37	224	970	818	19	44	40	
SET FILLS ONLY	543	191	89	17	23	5	16	62	414	463	18	51	32	0

Figure 1: Unisys MUC-3 System Scores

TEST RESULTS

The scores reported for the Unisys MUC-3 system are shown in Figure 1. The low ACT and high MIS scores reported for the *template id* slot indicate that event detection was a problem.[2] Poor event detection performance explains the relatively low recall scores reported for all but the *MATCHED ONLY* summary measurement. The *MATCHED ONLY* recall score is a measure of performance in which spurious (false positive) and missing (false negative) templates are not factored in. The extremely low SPU score reported for the *template id* slot suggests that further training of the rule base to improve event detection will not come at the expense of lower precision scores. In Figure 2, the performance of the Unisys system with respect to other MUC-3 systems is indicated in two scatter plots.

Since template slot-filling algorithms are triggered by the detection of an event, poor event detection performance has a direct negative impact on slot-filling performance. The recall scores for the Unisys MUC-3 system reflect this fact. However, for five slots precision scores are also low. These low precision scores are not a consequence of poor event detection, but result instead from a combination of poorly trained inference rules used to extract the sort of information expressed in the pertinent slots, and bugs in the template generation routines that gather and merge correctly detected information into template structures.

ANALYSIS

Contrary to what the low recall scores that have been reported suggest, the Unisys MUC-3 system can perform well at predicting events. The keyword-based prediction of event types is very robust; the database used during this stage of processing was derived from the full 1300 message DEV corpus. Moreover, when the rules used by KBIRD are properly trained, they do a very good job of locating instances of the events predicted by keyword analysis. Unfortunately, the KBIRD *locator* rules used to detect instances of events were trained on a relatively small set of messages—the 200 NOSC DEV and TST1 messages. Consequently, even though the keyword-based analysis phase may have correctly predicted the likely occurrence of a given event type, KBIRD may not have been able to locate an instance of the predicted event type. Thus, KBIRD's locator rules had a negating influence on the performance of the keyword-based analysis phase. Prior to the final MUC-3 test, versions of the Unisys system with fewer, more

[2]The *template id* slot is scored differently from other slots—the values reported for this slot are a measure of event detection performance (it doesn't make sense to report system performance in generating template ids, since the order in which templates are generated is not relevant in this task) [2].

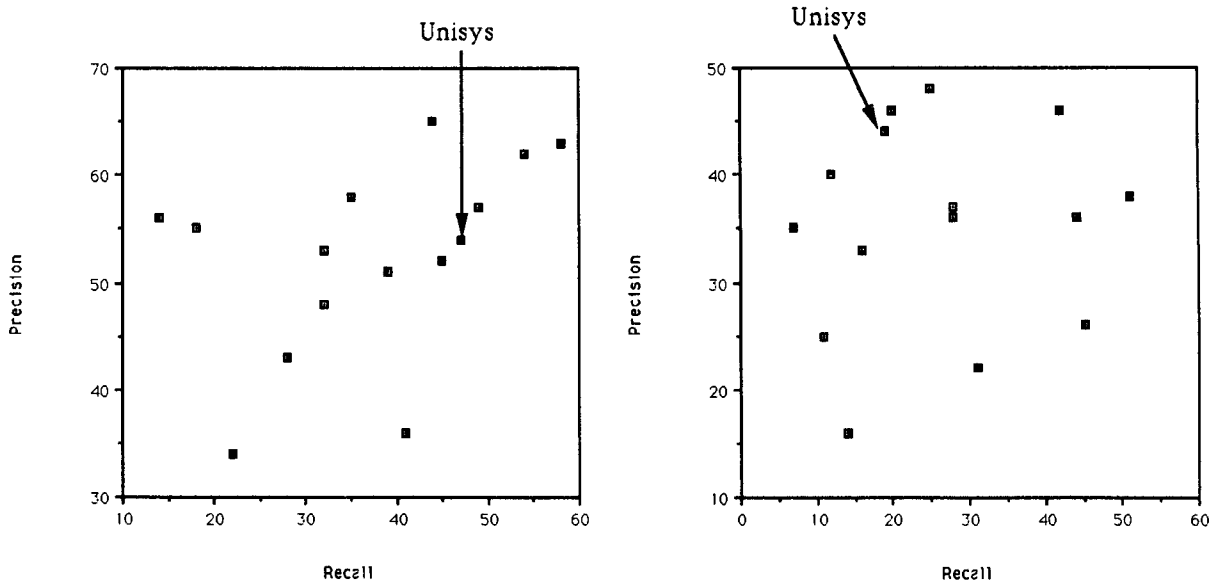

Figure 2: The scatter plot on the left indicates the relative performance of the Unisys MUC-3 system without taking into consideration false negative and false positive hits (the MATCHED-ONLY score). The scatter plot on the right indicates the relative performance of the Unisys MUC-3 system when taking into consideration both false negative and false positive hits (the ALL-TEMPLATES score).

general event detection rules in place had recall scores ranging in the high 30's and low 40's for all the summary measures. A tactical mistake was made in attempting to replace this general rule base with a larger, more context-sensitive one, since there was not enough time to allow the larger rule base to be properly trained. In the evaluation, generating spurious templates tended to have much less of an impact on scores than failing to generate templates at all. In future evaluations, we will investigate the use of different locator rule sets as a settable system parameter.

Rule training was hindered during the MUC-3 development cycle by the need to concurrently build the component that would be using the rules. In addition to this development problem, technical difficulties in KBIRD's design began to appear once the number of rules had grown to a realistic size. These technical problems resulted in slow message processing speeds, which further complicated the rule training process. The following three key problems were identified:

Heavy use of forward-chaining.
: There is currently too much reliance on forward-chaining in the KBIRD system. Many KBIRD reasoning tasks could be more efficiently achieved in a backward-chaining fashion.

Expensive TMS system.
: KBIRD was built on top of a very general inferencing mechanism with an expensive TMS system. KBIRD's needs for truth maintenance could be accomodated using a much simpler TMS component.

Inability to focus search.
: In KBIRD, it is currently not possible to focus search on a specific region of text. The mechanism used to satisfy a rule looks for all chart elements (concepts, words, phrases, and so forth) that match constituent expressions in the antecedent of a rule. If the KBIRD rule specifies that an element of a certain type must be in the same sentence as some other element, it would be more efficient to limit the search space to just those chart elements that fall within the span of the sentence. However, KBIRD's algorithm currently searches through chart elements indexed to locations anywhere in the text for suitable candidates.

CONCLUDING REMARKS

The time constraints imposed in MUC-3 made it impossible to fully develop the Unisys MUC-3 system's knowledge-based information retrieval component, KBIRD, before the evaluation deadline. Consequently, it is not possible at this time to establish the capabilities of the three-tiered approach realized in the system. The system's scores indicate, however, that although the rules for locating instances of events were inadequately trained, its performance at identifying slot values once an instance has been found is quite good.

Future work on the system will solve the technical problems that have been observed. This will be achieved by performing the following tasks:

- The overall system flow will be restructured to allow backward-chaining to handle more of the processing load.
- The current forward-chaining mechanism will be reimplemented so that it is specifically geared to the processing tasks envisoned for KBIRD.
- Subject to an appropriate funding source, the KBIRD *locator* rules used to detect instances of predicted event types will be properly trained.

In addition to solving the technical problems that have arisen in the system's KBIRD component, a major effort will be made to incorporate the Unisys Pundit NLP system into the MUC-3 system.

REFERENCES

[1] Tim Finin, Robin McEntire, Carl Weir, and Barry Silk. A three-tiered approach to natural language text retrieval. In *Proceedings of the AAAI workshop on Natural Language Text Retrieval*, Los Angeles, July 1991.

[2] NOSC. *Muc Scoring Manual*, 1991. Manual prepared for use in the Darpa-sponsored *Third Message Understanding Conference* (MUC-3).

[3] Carl Weir, Tim Finin, Barry Silk, Marcia Linebarger, and Robin McEntire. Knowledge-based strategies for robust text-understanding. The Eighth Annual Intelligence Community AI/Advance Computing Symposium, March 1991.