MCDONNELL DOUGLAS ELECTRONIC SYSTEMS COMPANY MUC-4 TEST RESULTS AND ANALYSIS

Amnon Meyers and David de Hilster

McDonnell Douglas Electronic Systems Company
Advanced Computing Technologies Lab
1801 E. St. Andrew Place
Santa Ana, CA 92705-6520
{vox, hilster}@young.mdc.com
(714) 566-5956

RESULTS

Our test results for TST3 and TST4 were as follows:

TST3		Recall	Precision	Overgeneration
	MATCHED/MISSING	20	60	13
	MATCHED/SPURIOUS	41	30	57
	MATCHED ONLY	41	60	13
	ALL TEMPLATES	20	30	57
	SET FILLS ONLY	24	69	14
	STRING FILLS ONLY	13	46	16
		P&R	2P&R	P&2R
	F-MEASURES	24.0	27.27	21.43

TST4		Recall	Precision	Overgeneration
	MATCHED/MISSING	23	63	11
	MATCHED/SPURIOUS	46	32	56
	MATCHED ONLY	46	63	11
	ALL TEMPLATES	23	32	56
	SET FILLS ONLY	26	70	12
	STRING FILLS ONLY	18	59	15
		P&R	2P&R	P&2R
	F-MEASURES	26.76	29.68	24.37

Our MUC4 TexUS system not only incorporates many of the concepts, knowledge, and algorithms from our MUC3 INLET system, but also represents our first version of a domain-independent end-to-end natural language processing capability. Although we are excited that our system correctly process text from the lexical to the discourse level for the first time, we stress that the MUC4 system is a framework that will not be complete until the end of the year. Relative to MUC3, our scores reflect an increase in precision due to our system's end-to-end completeness, and a lowering of recall due to the current incompleteness of our knowledge base and implementation. As TexUS nears completion, we expect our performance on the MUC4 task to improve substantially.

TEST SETTINGS

No test settings were incorporated during testing.

EFFORT

A total of 3.5 person-months were expended during our MUC4 customization effort. Five people were involved in MUC4 related activities: Amnon Meyers (task leader, analyzer, knowledge addition), David de Hilster (analyzer, knowledge addition, testing), Charlene Kowalski and Laurie Johnston (corpus study, knowledge addition), and George Vamos (automated testing).

LIMITING FACTORS

The greatest limiting factor for us again was time. Development of our new system TexUS reflects about one man-year of development and less than 4 man-months for the MUC4 task. Nevertheless, in a period of less than two months with a total of 3 persons, TexUS received a new and very thorough lexical processor and new baseline semantic and discourse processors. According to our development plans, the performance of the system should match that of the best participants at MUC4 by the end of 1992.

TRAINING THE SYSTEM

Extensive testing facilities were set up to automatically run and score TexUS on all 1300 messages of the development corpus. Testing was distributed on 5 to 9 Sun workstations (running in the background) processing 1300 messages within 6 to 10 hours. In addition, any 100 message set could be distributed, processed, and scored within one hour.

Another program for testing key phrases, linguistic phenomena, and slot fills, was also set up to monitor the consistency and progress of TexUS. These tests consist of pairs of small hand-made MUC4 templates and their corresponding key-templates. The current version of TexUS is then run on these hand-made templates and the differences printed out to monitor progress. One advantage of this testing format is that performance on individual slots is easy to isolate.

Due to a lack of time and incompleteness of the current system, the testing facilities were not extensively used prior to the official testing. The few times they were used on all 1300 messages of the development set, the results provided useful feedback. These tools, in conjunction with the development set, will be used extensively between MUC4 and MUC5 to monitor and speed development of our system.

SUCCESSES AND NON-SUCCESSES

Our greatest success was in the fact that our system is now an end-to-end system. Last year with only the pattern matching algorithm complete, we quickly reached the limitations of such a capability and our score could not go much higher without MUC-specific prodding. This year's system, TexUS, being a complete end-to-end system, has great potential to grow and score much higher than last year's system.

The lack of development time was the major cause of our difficulties. We needed more time to add knowledge and expand the linguistic coverage of the system.

REUSABILITY

One of TexUS' greatest strengths is its portability to new domains. TexUS not only is written in C and can be compiled into any C or C++ program, but it also is domain-independent. Only three customization tasks were needed to perform the MUC4 task: identification of key words and concepts, special semantic classification of words for the MUC4 domain, and a post-processing module to convert generic output to MUC4 output.

Less than 10 percent of the current code is dedicated to MUC4-like processing, 90 percent of which consists of template generation and merging as dictated by MUC4 template guidelines. The time needed to customize TexUS to a new domain is a matter of a few man-months, as demonstrated by our system during MUC3 and MUC4. Development time is cut further by the use of our graphic interface tools.

LESSONS LEARNED

- At MUC3, we fielded a system with limited potential, and substantially achieved that potential. At MUC4, we have fielded a system with far greater potential, but have not yet achieved that potential.

- Regarding the MUC task, we realize that greater manpower resources will be required to achieve the performance of top-ranking systems.

- As in MUC3, MUC4 demonstrates the importance of test and evaluation in driving development.