THIRD MESSAGE UNDERSTANDING EVALUATION AND CONFERENCE (MUC-3): PHASE 1 STATUS REPORT

Beth M. Sundheim

Naval Ocean Systems Center
Code 444
San Diego, CA 92152-5000

ABSTRACT

The Naval Ocean Systems Center is conducting the third in a series of evaluations of English text analysis systems. The premise on which the evaluations are based is that task-oriented tests enable straightforward comparisons among systems and provide useful quantitative data on the state of the art in text understanding. Furthermore, the data can be interpreted in light of information known about each system's text analysis techniques in order to yield qualitative insights into the relative validity of those techniques as applied to the general problem of information extraction. A dry-run phase of the third evaluation was completed in February, 1991, and the official testing will be done in May, 1991, concluding with the Third Message Understanding Conference (MUC-3). Twelve sites reported results for the dry-run test at a meeting held in February, 1991. All systems are being evaluated on the basis of performance on the information extraction task in a blind test at the end of each phase of the evaluation.

BACKGROUND

The Naval Ocean Systems Center (NOSC) is extending the scope of previous efforts in the area of evaluating English text analysis systems. These evaluations are intended to advance our understanding of the merits of current text analysis techniques, as applied to the performance of a realistic information extraction task. The current one is also intended to provide insight into information retrieval technology (document retrieval and categorization) used instead of or in concert with language understanding technology. The inputs to the analysis/extraction process consist of naturally-occurring texts that were obtained by NOSC in the form of electronic messages. The outputs of the process are a set of templates or semantic frames resembling the contents of a partially formatted database.

The premise on which the evaluations are based is that task-oriented tests enable straightforward comparisons among systems and provide useful quantitative data on the state of the art in text understanding. Even though the tests are designed to treat the systems under evaluation as black boxes, they are also designed to point up system performance on individual aspects of the task as well as on the task overall. Furthermore, these quantitative data can be interpreted in light of information known about each system's text analysis techniques in order to yield qualitative insights into the relative validity of those techniques as applied to the general problem of information extraction.

SCOPE

The third evaluation began in October, 1990. A dry-run phase was completed in February, 1991, and the official testing will be carried out in May, 1991, concluding with the Third Message Understanding Conference (MUC-3). This evaluation is significantly broader in scope than previous ones in most respects, including text characteristics, task specifications, performance measures, and range of text understanding and information extraction techniques. The corpus and task are sufficiently challenging that they are likely to be used again (with a new test set) in a future evaluation of the same and/or similar systems.

The corpus was formed via a keyword query to an electronic database containing articles in message format from open sources worldwide, compiled, translated (if necessary), edited, and disseminated by the Foreign Broadcast Information Service of the U.S. Government. A training set of 1300 texts was identified, and

additional texts were set aside for use as test data. The corpus presents realistic challenges in terms of its overall size (over 2.5 mb), the length of the individual articles (approximately half-page each on average), the variety of text types (newspaper articles, summary reports, speech and interview transcripts, rebel communiques, etc.), the range of linguistic phenomena represented (both well-formed and ill-formed), and the open-ended nature of the vocabulary (especially with respect to proper nouns).

TST1-MUC3-0080

BOGOTA, 3 APR 90 (INRAVISION TELEVISION CADENA 1) -- [REPORT] [JORGE ALONSO SIERRA VALENCIA] [TEXT] LIBERAL SENATOR FEDERICO ESTRADA VELEZ WAS KIDNAPPED ON 3 APRIL AT THE CORNER OF 60TH AND 48TH STREETS IN WESTERN MEDELLIN, ONLY 100 METERS FROM A METROPOLITAN POLICE CAI [IMMEDIATE ATTENTION CENTER]. THE ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER HAD LEFT HIS HOUSE WITHOUT ANY BODYGUARDS ONLY MINUTES EARLIER. AS HE WAITED FOR THE TRAFFIC LIGHT TO CHANGE, THREE HEAVILY ARMED MEN FORCED HIM TO GET OUT OF HIS CAR AND GET INTO A BLUE RENAULT.

HOURS LATER, THROUGH ANONYMOUS TELEPHONE CALLS TO THE METROPOLITAN POLICE AND TO THE MEDIA, THE EXTRADITABLES CLAIMED RESPONSIBILITY FOR THE KIDNAPPING. IN THE CALLS, THEY ANNOUNCED THAT THEY WILL RELEASE THE SENATOR WITH A NEW MESSAGE FOR THE NATIONAL GOVERNMENT.

LAST WEEK, FEDERICO ESTRADA VELEZ HAD REJECTED TALKS BETWEEN THE GOVERNMENT AND THE DRUG TRAFFICKERS.

Figure 1. Sample MUC-3 Terrorist Message

The task is to extract information on terrorist incidents (incident type, date, location, perpetrator, target, instrument, outcome, etc.) from the relevant messages in a blind test on 100 previously unseen texts in the test set. Approximately half of the messages will be irrelevant to the task as it has been defined. The extracted information is to be represented in the template in one of several ways, according to the information requirements of each slot. Some fills are required to be categories from a predefined set of possibilities (e.g., for the various types of terrorist incidents such as BOMBING, ATTEMPTED BOMBING, BOMB THREAT); others are required to be canonicalized forms (e.g., for dates) or numbers; still others are to be in the form of strings (e.g., for person names). The participants collectively created a set of training templates, each site manually filling in templates for 100 messages. A simple text and corresponding answer-key template are shown in Figures 1 and 2. Note that the text in Figure 1 is all upper case, that the dateline includes the source of the article ("Inravision Television Cadena 1") and that the article is a news report by Jorge Alonso Sierra Valencia.

0. MSG ID	TST1-MUC3-0080
1. TEMPLATE ID	1
2. INCIDENT DATE	03 APR 90
3. INCIDENT TYPE	KIDNAPPING
4. INCIDENT CATEGORY	TERRORIST ACT
5. INDIV PERPETRATORS	"THREE HEAVILY ARMED MEN"
6. ORG PERPETRATORS	"THE EXTRADITABLES" / "EXTRADITABLES"
7. PERP CONFIDENCE	REPORTED AS FACT: "THREE HEAVILY ARMED MEN" CLAIMED OR ADMITTED: "THE EXTRADITABLES" / "EXTRADITABLES"
8. PHYS TARGET ID	*
9. PHYS TARGET NUM	*
10. PHYS TARGET TYPE	*
11. HUM TARGET ID	"FEDERICO ESTRADA VELEZ" ("LIBERAL SENATOR" /"ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER" / "SENATOR" / "LIBERAL PARTY LEADER" / "PARTY LEADER")
12. HUM TARGET NUM	1
13. HUM TARGET TYPE	GOVERNMENT OFFICIAL / POLITICAL FIGURE
14. FOREIGN TGT NAT'N	-
15. INSTRUMENT TYPE	*
16. INCIDENT LOCATION	COLOMBIA: MEDELLIN (CITY)
17. PHYS TGT EFFECT	*
18. HUM TGT EFFECT	-

Figure 2. Sample Key Template

In Figure 2, the slot labels have been abbreviated to save space. The right-hand column contains the "correct answers" as defined by NOSC. Slashes mark alternative correct responses (systems are to generate just one of the possibilities), an asterisk marks slots that are inapplicable to the incident type being reported, and a hyphen marks a slot for which the text provides no fill.

A call for participation was sent to organizations in the U.S. that were known to be engaged in system design or development in the area of text analysis or information retrieval. Twelve of the sites that responded participated in the dry run and reported results at a meeting held in February, 1991. These sites are Advanced Decision Systems (Mountain View, CA), General Electric (Schenectady, NY), GTE (Mountain View, CA), Intelligent Text Processing, Inc. (Santa Monica, CA), Language Systems, Inc. (Woodland Hills, CA), New York University (New York City, NY), Planning Research Corp. (McLean, VA), SRI International (Menlo Park, CA), TRW (Redondo Beach, CA), Unisys CAIT (Paoli, PA), the University of Massachusetts (Amherst, MA), and the University of Nebraska (Lincoln, NE) in association with the University of Southwest Louisiana (Lafayette, LA). The meeting also served as a forum for resolving issues that affect the test design, scoring, etc. for the official test in May.

A wide range of text interpretation techniques (e.g., statistical, key-word, template-driven, pattern-matching, and natural language processing) were represented in this phase of the evaluation. One of the participating sites, TRW, offered a preliminary baseline performance measure for a pattern-matching approach to information extraction that they have already successfully put into operational use as an interactive system applied to texts of a somewhat more homogeneous and straightforward nature than those found in the MUC-3 corpus. All sites reporting in February are likely to continue development in phase 2 and undergo official testing in May. In addition, three sites that did not report results for the dry run are expecting to report results on the official run.

MEASURES OF PERFORMANCE

All systems are being evaluated on the basis of performance on the information extraction task in a blind test at the end of each phase of the evaluation. It is expected that the degree of success achieved by the different techniques in May will depend on such factors as whether the number of possible slot fillers is small, finite, or open-ended and whether the slot can typically be filled by fairly straightforward extraction or not. System characteristics such as amount of domain coverage, degree of robustness, and general ability to make proper use of information found in novel input will also be major factors. The dry-run test results cannot be assumed to provide a good basis for estimating performance on the official test in May.

An excellent, semi-automated scoring program has been developed and distributed to all participants to enable the calculation of the various measures of performance. The two primary measures are completeness (recall) and accuracy (precision). There are two additional measures, one to isolate the amount of spurious data generated (overgeneration) and the other to determine the rate of incorrect generation as a function of the number of opportunities to incorrectly generate (fallout). Fallout can be calculated only for those slots whose fillers form a closed set. Scores for the other three measures are calculated for the test set overall, with breakdowns by template slot. Figure 3 presents a somewhat simplified set of definitions for the measures.

MEASURE	DEFINITION
RECALL	$\frac{\text{\#correct fills generated}}{\text{\#fills in key}}$
PRECISION	$\frac{\text{\#correct fills generated}}{\text{\#fills generated}^{*}}$
OVER-GENERATION	$\frac{\text{\#spurious fills generated}}{\text{\#fills generated}}$
FALLOUT	$\frac{\text{\#incorrect+spurious gen'ed}}{\text{\#possible incorrect fills}}$

Figure 3. MUC-3 Scoring Metrics

The most significant things to note are that precision and recall are actually calculated on the basis of points -- the term "correct" includes system responses that matched the key exactly (earning 1 point each) and system responses that were judged to be a good partial match (earning .5 point each). It should also be noted that overgeneration figures in precision by contributing to the denominator in addition to being isolated as a measure in its own right. Overgeneration also figures in fallout by contributing to the numerator. This fact will come up again in the next section in the discussion of the phase 1 results.

In addition to the official measures, unofficial measures will be obtained in May of performance on particular linguistic phenomena (e.g., conjunction), as measured by the database fills generated by the systems in particular sets of instances. That is, sentences exemplifying a selected phenomenon will be marked for separate scoring if successful handling of the phenomenon seems to be required in order to fill one or more template slots correctly for that sentence. An experiment involving several phenomena tests was conducted as part of the dry run. The tests concerned the interpretation of active versus passive clauses, main versus embedded clauses, conjunction of noun phrases, and negation. The results for the dry run were extremely inconclusive, given the lack of basic domain coverage of the systems and, for several sites, the exclusive use of nonlinguistic processing components. In addition, the utility of this means of judging linguistic coverage was eroded by the fact that most systems had multiple points of failure; some may have handled the linguistic phenomena correctly in the early stages of analysis, but failed to fill the slots correctly due to subsequent processing failure.

PHASE 1 RESULTS

The results obtained in the first phase of the evaluation are unofficial and will therefore not be presented in their entirety. To give readers an idea of the current top level of performance of the participating systems, scores from two systems are presented anonymously. Table 1 presents a summary of the scores obtained on recall, precision, and overgeneration for the system that scored highest overall on recall and the system that scored highest overall on precision (with recall above a threshold of 10%). The results for the fallout measure cannot be calculated for the test overall (because the fillers for some slots do not form closed sets) and are therefore not included in Table 1.

SELECTION CRITERION	MEASURE		
	RCL	PRC	OVG
S1: SYSTEM W/HIGHEST RECALL	**52**	60	22
S2: SYSTEM W/HIGHEST PRECISION	14	**68**	11

Table 1. Summary of Phase 1 Scores (%) for Best-Performing Systems on Recall and Precision

SLOT	RCL		PRC		OVG		FLT	
	S1	S2	S1	S2	S1	S2	S1	S2
1	87	38	56	74	43	26		
2	64	12	78	62	00	00		
3	81	34	93	91	00	00	00	00
4	50	34	70	62	19	18	06	11
5	25	06	55	50	25	22		
6	22	00	93		00			
7	28	09	58	83	08	11	04	00
8	74	03	55	17	36	11		
9	55	13	58	56	10	11		
10	43	02	35	50	32	00	03	00
11	50	10	55	50	26	00		
12	49	18	58	68	09	00		
13	41	22	41	94	29	00	05	00
14	26	00	59		27		00	00
15	17	00	23		62		01	00
16	63	00	72	25	00	00		
17	57	03	78	50	10	50	01	00
18	41	00	34		53		05	00

Table 2. Breakdown of Phase 1 Scores (%) by Template Slot for the Best-Performing Systems on Recall and Precision

Systems will tend to show a performance trade-off between recall and precision. S1 has

nearly four times greater recall than S2, and so it is not surprising that its overgeneration score is significantly worse than S2's. In this regard, it should be noted that generating a spurious template incurs a penalty that affects only slot 1, the template ID slot. Thus, although the precision of S1 is lower than S2's as expected, the difference is not nearly as marked as it would be if the penalty for generating a spurious template affected all slots rather than just the template ID slot.

The recall columns in Table 2 suggest to what extent S1 and S2 have been developed to fill data in the various template slots. S2 has zero percent recall for several of the slots. In the particular case of S2, a system based on thorough syntactic and semantic analysis, the reason for the zero reall is that system development simply has not focused yet on filling those slots. Only one (slot 4) requires a string fill; the other three take a set fill. However, in the case of systems based on text categorization techniques (not represented in Tables 1 and 2), zero recall is more likely to appear consistently in the slots whose fillers do not form a closed set, reflecting an inherent limitation in the approach. In order to obtain measures that give a fair appraisal of all systems in terms of their ability to select proper categories of responses, it has been suggested that a second set of "overall" measures be calculated that includes only those slots for which the fillers form a closed set.

As defined for MUC-3, the numerator for fallout includes both the number of spurious slot fillers and the number of incorrect slot fillers. The inclusion of the spurious fillers in the numerator changes the intended meaning of the measure, as seen in the results for slot 4 in Table 2. That slot can be filled with one of only two possible set fills, either STATE-SPONSORED VIOLENCE or TERRORIST ACT, or it sometimes is intended to be null (represented as a hyphen in the notation). All other slots for which fallout can be computed have significantly more options, i.e., "opportunities to incorrectly generate." If the fallout score were computed without including spurious fillers, the scores for the CATEGORY OF INCIDENT slot should be relatively low compared to the other slot scores for fallout. Instead, the scores for fallout on that slot are higher than for any of the others, probably showing that the systems frequently filled that slot when it was supposed to be null.

ACKNOWLEDGEMENTS

The author is indebted to all the organizations participating in MUC-3 and to certain individuals in particular who have contributed extra time and energy to ensure the evaluation's success, among them Laura Balcom, Sean Boisen, Nancy Chinchor, Ralph Grishman, Pete Halverson, Jerry Hobbs, Cheryl Kariya, George Krupka, David Lewis, Lisa Rau, John Sterling, Charles Wayne, and Carl Weir.