OVERVIEW OF THE THIRD MESSAGE UNDERSTANDING EVALUATION AND CONFERENCE

Beth M. Sundheim

Naval Ocean Systems Center
Code 444
Decision Support and AI Technology Branch
San Diego, CA 92152-5000
sundheim@nosc.mil

INTRODUCTION

The Naval Ocean Systems Center (NOSC) has conducted the third in a series of evaluations of 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 latest 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 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 these 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. The tests are designed to treat the systems under evaluation as black boxes and to point up system performance on discrete aspects of the task as well as on the task overall. 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.

The process of conducting these evaluations has presented great opportunities for examining and improving on the evaluation methodology itself. Although still far from perfect, the MUC-3 evaluation was markedly better than the previous one, especially with respect to the way scoring was done and the degree to which the test set was representative of the training set. Much of the credit for improvement goes to the evaluation participants themselves, who have been actively involved in nearly every aspect of the evaluation. The previous MUC, known as MUCK-II (the naming convention has since been stripped down), proved that systems existed that could do a reasonable job of extracting data from ill-formed paragraph-length texts in a narrow domain (naval messages about encounters with hostile forces) and that measuring performance on such a task was a feasible and viable thing to do. However, the usage of a very small test set (just 5 texts) and an extremely unsophisticated scoring procedure combined to make it inadvisable to publicize the

results. (Results obtained in experiments conducted on one MUCK-II system after the evaluation was completed are discussed in [1].)

The MUC-3 evaluation was 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. MUC-3 presented a significantly more challenging task than MUCK-II, which was held in June of 1989. The results show that MUC-3 was not an unreasonable challenge to 1991 technologies. The means used to measure performance have evolved far enough that we no longer hesitate to present the system scores, and work on the evaluation methodology is planned that will take the next step to determine the statistical significance of the results.

In another effort to determine their significance, some work has already been undertaken by Hirschman [2] to measure the difference in complexity of MUC-like evaluation tasks so that the results can be used to quantify progress in the field of text understanding. This objective, however, brings up another critical area of improvement for future evaluations, namely refining the evaluation methodology in such a way as to better isolate the systems' text analysis capabilities from their data extraction capabilities. This will be done, since the MUC- 3 corpus and task are sufficiently challenging that they can be used again (with a new test set) in a future evaluation. That evaluation will seek to examine more closely the text analysis capabilities of the systems, to measure improvements in performance by MUC-3 systems, and to establish performance baselines for any new systems.

This paper covers most of the basics of the MUC-3 evaluation, which were presented during a tutorial session and in an overview presentation at the start of the regular sessions. This paper is also an overview of the conference proceedings, which includes papers contributed by the sites that participated in the evaluation and by individuals who were involved in the evaluation in other ways. Parts I, II, and III of the proceedings are organized in the order in which the sessions were held, but the ordering of papers within Parts II and III is alphabetical by site and does not necessarily correspond with the order in which the presentations were made during the conference. The proceedings also includes a number of appendices containing materials pertinent to the evaluation.

OVERVIEW OF MUC-3

The planning for MUC-3 began while MUCK-II was still in progress, with suggestions from MUCK-II participants for improvements. A MUC-3 program committee was formed from among those MUCK-II participants who provided significant feedback on the MUCK-II effort. The MUC-3 program committee included Laura Blumer Balcom (Advanced Decision Systems), Ralph Grishman (New York University), Jerry Hobbs (SRI International), Lisa Rau (General Electric), and Carl Weir (Unisys Center for Advanced Information Technology). Since one of the suggestions for MUC-3 was to add an element of document filtering to the task of data extraction, David Lewis (then at the University of Massachusetts and now at the University of Chicago) was invited to join the committee as a representative of the information retrieval community.

NOSC began looking for a suitable corpus in late 1989 and obtained assistance from other government agencies to acquire it during the summer of 1990. At that

time, a call for participation was sent to academic, industrial, and commercial organizations in the United States that were known to be engaged in system design or development in the area of text analysis or information retrieval. Participation on the part of many of the respondents was contingent upon receiving outside financial support; approximately two-thirds of the sites were awarded financial support by the Defense Advanced Research Projects Agency (DARPA). These awards were modest, some sites having requested funds only to pay travel expenses and others having requested funds to cover up to half of the total cost of participating. The total cost was typically estimated to be approximately equivalent to one person-year of effort.

The evaluation was officially launched in October, 1990, with a three-month phase dedicated to compiling the "answer key" templates for the texts in the training set (see next section), refining the task definition, and developing the initial MUC-3 version of the data extraction systems. These systems underwent a dry-run test in February, 1991, after which a meeting was held to discuss the results and hammer out some of the remaining evaluation issues. Twelve sites participated in the dry run. One site dropped out after the dry run (TRW), and four new sites entered, three of which had already been involved to some extent (BBN Systems and Technologies, McDonnell Douglas Electronic Systems Company, and Synchronetics, Inc.) and one that had not (Hughes Research Laboratories).

The second phase began in mid-February and, while system development continued at each of the participating sites, updates were made to the scoring program, the task definition, and the answer key templates for the training set. Final testing was carried out in May, 1991, concluding with the Third Message Understanding Conference (MUC-3), which was attended by representatives of the participating sites and interested government organizations. During the conference, the evaluation participants decided that the test results should be validated by having the system-generated templates rescored by a single party. Two of the participants were selected to work as a team to carry out this task, and the results of their effort are the official test scores presented in this volume.

Pure and hybrid systems based on a wide range of text interpretation techniques (e.g., statistical, key-word, template-driven, pattern-matching, in-depth natural language processing) were represented in the MUC-3 evaluation. The fifteen sites that completed the evaluation are Advanced Decision Systems (Mountain View, CA), BBN Systems and Technologies (Cambridge, MA), General Electric (Schenectady, NY), General Telephone and Electronics (Mountain View, CA), Intelligent Text Processing, Inc. (Santa Monica, CA), Hughes Research Laboratories (Malibu, CA), Language Systems, Inc. (Woodland Hills, CA), McDonnell Douglas Electronic Systems (Santa Ana, CA), New York University (New York City, NY), PRC, Inc. (McLean, VA), SRI International (Menlo Park, CA), Synchronetics, Inc. together with the University of Maryland (Baltimore, MD), Unisys Center for Advanced Information Technology (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). Parts II and III of this volume include papers by each of these sites. In addition, an experimental prototype of a probabilistic text categorization system was developed by David Lewis, who is now at the University of Chicago, and was tested along with the other systems. That work is described in a paper in Part IV.

CORPUS AND TASK

The corpus was formed via a keyword query[1] to an electronic database containing articles in message format from open sources worldwide. These articles had been gathered, translated (if necessary), edited, and disseminated by the Foreign Broadcast Information Service (FBIS) of the U.S. Government. A training set of 1300 texts was identified, and additional texts were set aside for use as test data[2]. The message headers were used to create or augment a dateline and the text type information appearing at the front of the article; the original message headers and routing information were removed. The layout was modified slightly to improve readability (e.g., by double-spacing between paragraphs), and problems that arose with certain characters when the data was downloaded were rectified (e.g., square brackets were missing and had to be reinserted). The body of the text was modified minimally and with the sole purpose of eliminating some idiosyncratic features that were well beyond the scope of interest of MUC-3[3].

The corpus presents realistic challenges in terms of overall size (over 2.5 megabytes), length of the individual articles (approximately a half-page each on average), variety of text types (newspaper articles, TV and radio news, speech and interview transcripts, rebel communiques, etc.), range of linguistic phenomena represented (both well-formed and ill-formed), and open-endedness of the vocabulary (especially with respect to proper nouns). The texts used in MUCK-I and MUCK-II originated as teletype messages and thus were all upper case; the MUC-3 texts are also all upper case, but only as a consequence of downloading from the source database, where the texts appear in mixed upper and lower case.

The task was to extract information on terrorist incidents (incident type, date, location, perpetrator, target, instrument, outcome, etc.) from the relevant texts in a blind test on 100 previously unseen texts. Approximately half the articles were irrelevant to the task as defined. In some cases the terrorism keywords in the query used to form the corpus (see footnote) were used in irrelevant senses, e.g., "explosion" in the phrase "social explosion". In other cases, an entity of one of the

---

[1]The query specified a hit as a message containing both a country/nationality name (e.g., Honduras or Honduran) for one of the nine countries of interest (Argentina, Bolivia, Chile, Colombia, Ecuador, El Salvador, Guatemala, Honduras, Peru) and some inflectional form of a common word associated with terrorist acts (abduct, abduction, ambush, arson, assassinate, assassination, assault, blow [up], bomb, bombing, explode, explosion, hijack, hijacking, kidnap, kidnapping, kill, killing, murder, rob, shoot, shooting, steal, terrorist). Some of the articles in the MUC-3 corpus may no longer satisfy this query, since the message headers (including the subject line) were removed after the retrieval was done.

[2]Over 300 articles were set aside from the overall corpus to be used as test data. The composition of the test sets was intentionally controlled with respect to the frequency with which incidents concerning any given country are represented; otherwise, the selection was done simply by taking every *n*th article about that country.

[3]For example, transcriptions of radio and TV broadcasts sometimes contained sentences in which words were enclosed in parentheses to indicate that the transcriber could not be certain of them, e.g., "They are trying to implicate the (Ochaski Company) with narcoterrorism." (This quote is from article number PA1807130691 of the Latin America volume of the Foreign Broadcast Information Service Daily Reports.) In cases such as this, where the text is parenthetical in form but not in function, the parentheses were deleted.

nine countries of interest -- the second necessary condition for a hit -- was mentioned, but the entity did not play a significant role in the terrorist incident.

Other articles were irrelevant for reasons that were harder to formulate. For example, some articles concerned common criminal activity or guerrilla warfare (or other military conflict). Rules were developed to challenge the systems to discriminate among various kinds of violent acts and to generate templates only for those that would be of interest to a terrorism news analyst. The real-life scenario also required that only timely, substantive information be extracted; thus, rules were formulated that defined relevance in terms of whether the news was recent and whether it at least mentioned who/what the target was. Other relevance criteria were developed as well, again with the intent of simulating a real-life task. The relevance criteria are described in the first part of appendix A, which is the principal documentation of the MUC-3 task. Appendix D contains some representative samples of relevant and irrelevant articles.

It can be seen that the relevance criteria are extensive and would sometimes be difficult to state, let alone implement. It was learned that greater allowances needed to made for the fact that this was an evaluation task and *not* a real-life one. Systems that generated generally correct internal data structures for a relevant incident, only to filter out that data structure by making a single mistake on one of the relevance criteria, were penalized for having missed the incident entirely rather than being penalized for getting just one aspect of the incident description wrong. Some allowance *was* made in the answer key for the fact that incidents or facts about incidents might be of questionable relevance, given the vagueness of some texts and gaps in the statement of the relevance criteria; the template notation allowed for optionality, and systems were not penalized if they failed to generate an optional template or an optional filler in a required template.

If an article was determined to be relevant, there was then the task of determining how many distinct relevant incidents were being reported. The information on these incidents had to be correctly disentangled and represented in separate templates. The extracted information was to be represented in the template in one of several ways, according to the data format requirements of each slot. (See appendix A.) Some slot fills were required to be categories from a predefined set of possibilities called a "set list" (e.g., for the various types of terrorist incidents such as **BOMBING, ATTEMPTED BOMBING, BOMB THREAT**); others were required to be canonicalized forms (e.g., for dates) or numbers; still others were to be in the form of strings (e.g., for person names).

A relatively simple article and corresponding answer key template from the dry-run test set (labeled TST1) 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. In Figure 2, the left-hand column contains the slot labels, and 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, a hyphen marks a slot for which the text provides no fill, and a colon introduces the cross-reference portion of a fill (except for slot 16, where the colon is used as a separator between more general and more specific place names). More information on the template notation can be found in appendix A, and further examples of texts and templates can be found in appendices D and E.

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. Article from MUC-3 Corpus[4]

0. MESSAGE ID	TST1-MUC3-0080
1. TEMPLATE ID	1
2. DATE OF INCIDENT	03 APR 90
3. TYPE OF INCIDENT	KIDNAPPING
4. CATEGORY OF INCIDENT	TERRORIST ACT
5. PERPETRATOR: ID OF INDIV(S)	"THREE HEAVILY ARMED MEN"
6. PERPETRATOR: ID OF ORG(S)	"THE EXTRADITABLES" / "EXTRADITABLES"
7. PERPETRATOR: CONFIDENCE	CLAIMED OR ADMITTED: "THE EXTRADITABLES" / "EXTRADITABLES"
8. PHYSICAL TARGET: ID(S)	*
9. PHYSICAL TARGET: TOTAL NUM	*
10. PHYSICAL TARGET: TYPE(S)	*
11. HUMAN TARGET: ID(S)	"FEDERICO ESTRADA VELEZ" ("LIBERAL SENATOR" / "ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER" / "SENATOR" / "LIBERAL PARTY LEADER" / "PARTY LEADER")
12. HUMAN TARGET: TOTAL NUM	1
13. HUMAN TARGET: TYPE(S)	GOVERNMENT OFFICIAL / POLITICAL FIGURE: "FEDERICO ESTRADA VELEZ"
14. TARGET: FOREIGN NATION(S)	-
15. INSTRUMENT: TYPE(S)	*
16. LOCATION OF INCIDENT	COLOMBIA: MEDELLIN (CITY)
17. EFFECT ON PHYSICAL TARGET(S)	*
18. EFFECT ON HUMAN TARGET(S)	-

Figure 2. Answer Key Template

[4]This article has serial number PA0404072690 in the Latin America volume of the FBIS Daily Reports, which are the secondary source for all the texts in the MUC-3 corpus.

The participants collectively created the answer key for the training set, each site manually filling in templates for partially overlapping subset of the texts. This task was carried out at the start of the evaluation; it therefore provided participants with good training on the task requirements and provided NOSC with good early feedback. Generating and cross-checking the templates required an investment of at least two person-weeks of effort per site. These answer keys were updated a number of times to reduce errors and to maintain currency with changing template fill specifications. In addition to generating answer key templates, sites were also responsible for compiling a list of the place names that appeared in their set of texts; NOSC then merged these lists to create the set lists for the **TARGET: FOREIGN NATION** slot and **LOCATION OF INCIDENT** slot.

MEASURES OF PERFORMANCE

All systems were 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 was expected that the degree of success achieved by the different techniques in May would depend on such factors as whether the number of possible slot fillers was small, finite, or open-ended and whether the slot could 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 were also expected to be major factors. The dry-run test results were not assumed to provide a good basis for estimating performance on the final test in May, but the expectation was that most, if not all, of the systems that participated in the dry run would show dramatic improvements in performance. The test results show that some of these expectations were borne out, while others were not or were less significant than expected.

A semi-automated scoring program was developed under contract for MUC-3 to enable the calculation of the various measures of performance. It was distributed to participants early on during the evaluation and proved invaluable in providing them with the performance feedback necessary to prioritize and reprioritize their development efforts as they went along. The scoring program can be set up to score all the templates that the system generates or to score subsets of templates/slots. User interaction is required only to determine whether a mismatch between the system-generated templates and the answer key templates should be judged completely or partially correct. (A partially correct filler for slot 11 in Figure 2 might be **"VELEZ" ("LEADER")**, and a partially correct filler for slot 16 would be simply **COLOMBIA**.) An extensive set of interactive scoring guidelines was developed to standardize the interactive scoring. These guidelines are contained in appendix C. The scoring program maintains a log of interactions that can be used in later scoring runs and augmented by the user as the system is updated and the system-generated templates change.

The two primary measures of performance were completeness (recall) and accuracy (precision). There were 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). The labels "recall," "precision," and "fallout" were borrowed from the field of information retrieval, but the definitions of those terms had to be substantially modified to suit the template-generation task. The overgeneration metric has no correlate in the information retrieval field, i.e., a

MUC-3 system can *generate* indefinitely more data than is actually called for, but an information retrieval system cannot *retrieve* more than the total number of items (e.g., documents) that are actually present in the corpus.

Fallout can be calculated only for those slots whose fillers form a closed set. Scores for the other three measures were calculated for the test set overall, with breakdowns by template slot. Figure 3 presents somewhat simplified definitions.

MEASURE	DEFINITION
RECALL	$\frac{\text{\#correct fills generated}}{\text{\#fills in key}}$
PRECISION	$\frac{\text{\#correct fills generated}}{\text{\#fills generated}}$
OVERGENERATION	$\frac{\text{\#spurious fills generated}}{\text{\#fills generated}}$
FALLOUT	$\frac{\text{\#incorrect+spurious generated}}{\text{\#possible incorrect fills}}$

Figure 3. MUC-3 Scoring Metrics

The most significant thing that this table does not show is 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 is not only a measure in its own right but is also a component of precision, where it acts as a penalty by contributing to the denominator. Overgeneration also figures in fallout by contributing to the numerator. Further information on the MUC-3 evaluation metrics and scoring methods, including information on three different ways penalties for missing and spurious data were assigned, can be found elsewhere in this volume in the paper on evaluation metrics by Nancy Chinchor [3].

TEST PROCEDURE

Final testing was done on a test set of 100 previously unseen texts that were representative of the corpus as a whole. Participants were asked to copy the test package electronically to their own sites when they were ready to begin testing. Appendix B contains a copy of the test procedure. The testing had to be conducted and the results submitted within a week of the date when the test package was made available for electronic transfer. Each site submitted their system-generated templates, the outputs of the scoring program (score reports and the interactive scoring history file), and a trace of the system's processing (whatever type of trace the system normally produces that could serve to help validate the system's outputs). Initial scoring was done at the individual sites, with someone designated as interactive scorer who preferably had not been part of the system development team. After the conference, the system-generated templates for all sites were labeled anonymously and rescored by two volunteers in order to ensure that the official scores were obtained as consistently as possible.

The system at each site was to be frozen before the test package was transferred; no updates were permitted to the system until testing and scoring

were completed. Furthermore, no backing up was permitted during testing in the event of a system error. In such a situation, processing was to be aborted and restarted with the next text. A few sites encountered unforeseen system problems that were easily pinpointed and fixed. They reported unofficial, revised test results at the conference that were generally similar to the official test results and do not alter the overall picture of the current state of the art.

The basic test called for systems to be set up to generate templates that produced the "maximum tradeoff" between recall and precision, i.e., templates that achieved scores as high as possible and as similar as possible on both recall and precision. This was the normal mode of operation for most systems and for many was the *only* mode of operation that the developers had tried. Those sites that *could* offer alternative tradeoffs were invited to do so, provided they notified NOSC in advance of the particular setups they intended to test on.

In addition to the scores obtained for these metrics on the basic template-generation task, scores were obtained of system performance on the linguistic phenomenon of apposition, as measured by the template fills generated by the systems in particular sets of instances. That is, sentences exemplifying apposition were marked for separate scoring if successful handling of the phenomenon seemed to be required in order to fill one or more template slots correctly for that sentence. This test was conducted as an experiment and is described in the paper by Nancy Chinchor on linguistic phenomena testing [4].

TEST RESULTS AND DISCUSSION

The summary score reports produced for the tested systems by the scoring program are found in appendix F; scatter plots for selected portions of the final test results are shown in appendix G. Most of the figures in appendix G plot recall versus precision; a couple plot recall vs overgeneration, since the generation of spurious data is a significant element of precision with respect to a template generation task. The plots facilitate consideration of questions such as the following:

* On which aspect of the task (slot in the template) were the systems as a group most successful?

* How well did the systems handle time expressions (**DATE OF INCIDENT** slot)?

* How did the front-running systems on the overall measures differ with respect to individual slot performance?

* To what extent do the different ways of computing the scores (Matched/Missing, Matched Only, All Templates, and Set Fills Only) change the picture?

* To what extent was generation of spurious data taking place?

* To what extent did the individual systems' recall and precision represent tradeoffs?

Not included in the appendices are the detailed score reports produced by the scoring program for each of the system-generated templates. These reports permit consideration of other interesting questions such as how systems performed from one terrorist incident type to another and how they performed when a message contained more than one relevant incident report. It is also possible to use them together with the corresponding texts to answer questions such as how well systems handled newspaper articles versus TV and radio news reports and how well they handled incident reports that were spread out over a paragraph or across paragraphs rather than being completely described in a single sentence.

The appendices also do not include the results of a minor study of human performance on the MUC-3 final test. This study was conducted using two MUC-3 evaluators as subjects and measuring their performance individually compared to the official answer key, which was created by merging and correcting their individual draft keys. The evaluator with the lower scores for Matched/Missing had 87% recall, 91% precision, and 5% overgeneration. Needless to say, since these subjects were responsible for preparing the official answer key, their performance on the draft keys was undoubtedly higher than could be expected even from other highly trained persons. Another reason they are higher than would be obtained in a different study is that the two evaluators prepared the draft keys in two stages and reconciled most of the differences that arose in the first stage before starting the second stage. In the first stage, the evaluators identified which articles were relevant, how many templates would be generated for the relevant ones, and which incident types would be represented in each of the templates. In the second stage, the evaluators filled in the templates, with the assistance of an interactive software tool that provides some integrity checking, automatic fill-in, etc.

The plots in appendix G present an interesting picture of the MUC-3 results as a whole, but the significance of the numbers for each of the tested systems needs to be assessed on the basis of a careful reading of the papers in this volume that were submitted by each of the sites. To facilitate interpretation of the test results, the sites were asked to focus on the test scores and the evaluation experience in the first of those papers and to elaborate in their second paper on how the system -- *as it was actually implemented for MUC-3* -- works in general and how it is designed to handle the kinds of phenomena found in the MUC-3 corpus.

The level of effort that could be afforded by each of the sites varied considerably, as did the maturity of the systems at the start of the evaluation. All sites were operating under time constraints imposed by the evaluation schedule. In addition, the evaluation demands were a consequence of the intricacies of the task and of general corpus characteristics such as the following:

* The texts that are relevant to the MUC-3 task (comprising approximately 50% of the total corpus) are likely to contain more than one relevant incident.

* The information on a relevant incident may be dispersed throughout the text and may be intertwined with accounts of other (relevant or irrelevant) incidents.

* The corpus includes a mixture of material (newspaper articles, TV news, speeches, interviews, propaganda, etc.) with varying text structures and styles.

The scoring program produces four sets of overall scores, three of which are based on different means of assessing penalties for missing and spurious data. These sets of scores appear in the rows at the bottom of the score reports. The set called Matched/Missing is a compromise between Matched Only (which is more lenient than Matched/Missing) and All Templates (more stringent) and is used as the official one for reporting purposes. Figure G1 is based on the Matched/Missing method of assessing penalties. The fourth method does the scoring only for those slots that require set fills, i.e., fills that come from predefined sets of categories. Figure G4 is based on that method of scoring. The various methods are described more fully in [3].

The remainder of this section is a discussion of just a few of the figures in appendix G. (The data points in appendix G are labeled with abbreviated names of the 15 sites, and optional test runs are marked with the site's name and an "O" extension.) Figure Gl gives the most general picture of the results of MUC-3 final testing. It shows that precision always exceeds recall and that the systems with relatively high recall are also the ones that have relatively high precision. The latter fact inspires an optimistic attitude toward the promise of at least some of the techniques employed by today's systems -- further efforts to enhance existing techniques and extend the systems' domain coverage may lead to significantly improved performance on both measures. However, since all systems show better precision than recall, it appears that it will be a bigger challenge to obtain very high recall than it will be to achieve higher precision at recall levels that are similar to those achievable today. This observation hold true even for Figure G2 (Matched Only), where recall is substantially greater for most systems compared to G1.[5]

The distribution of data points tentatively supports at least one general observation about the technologies underlying today's systems: those systems that use purely stochastic techniques or handcrafted pattern-matching techniques were not able to achieve the same level of performance for MUC-3 as some of the systems that used parsing techniques. The "non-parsing" systems are ADS, HU, MDC, UNI, UNL, UNL-O1, and UNL-O2, and the "parsing" systems are BBN, BBN-O, GE, GTE, ITP, LSI, NYU, NYU-O1, NYU-O2, PRC, SRI, SYN, UMA, and UMA-O.

Further support for this observation can be found in Figure G4, where the scores are computed for all slots requiring set fills, and in Figure G9, which shows the scores for just one of those set-fill slots, the **TYPE OF INCIDENT**. In these cases, one might expect the non-parsing systems to compare more favorably with the parsing systems, since the fill options are restricted to a fairly small, predefined set of possibilities.[6] However, none of the non-parsing systems appears at the leading edge in Figure G4, and the only non-parsing system in the cluster at the leading edge in Figure G9 is ADS (which shares a data point with NYU-O2),

---

[5] Recall is greater in G2 because Matched Only differs from Matched/Missing in that the "total possible," i.e., the recall denominator, does not include penalties for missing templates.

[6] The results in G4 are somewhat contaminated due to the fact that some of the set-fill slots require that the fillers be cross-referenced to fillers of string-fill slots (see, for example, the fillers of slots 7, 11, and 13 in Figure 2 earlier in this paper). The scoring of the set-fill slots is affected by these cross-reference tags. However, the **TYPE OF INCIDENT** results (G9) are not contaminated in this way.

although a few non-parsing systems have extremely high precision scores (UNI, UNL, UNL-O1, and UNL-O2).

On the other hand, there is quite a range in performance even among the systems in the parsing group, all of which had to cope with having limited coverage of the domain. One thing that is apparent from the sites' system descriptions (see Part III of this proceedings) is that the ones on the leading edge in Figure G1 have the ability to make good use of partial sentence parses when complete parses cannot be obtained. Level of effort is also an indicator of performance success, though not a completely reliable one: GE, NYU, and UMass all reported investing more than one person-year of effort in MUC-3, but several other sites with lower overall performance also reported just under or over one person-year of effort.

It must be said that there were some extremely immature systems in the non-parsing group and the parsing group alike, so any general conclusions must be taken as tentative and should certainly not be used to form opinions about the relative validity of isolated techniques employed by the individual systems in each group. It could be that the relatively low-performing systems use extremely effective techniques that, if supplemented by other known techniques or supported by more extensive domain coverage, would put the system well out in front. Neither should one assume that the systems at the leading edge are similar kinds of systems. In fact, those systems have quite different architectures and have varying sizes of lexicons, kinds of parsers and semantic interpreters, etc.

Figures G7 through G24 show how system performance varied from one slot to another. Figures G7, G9, and G17 are useful as examples of the way spurious data generation combines with incorrect data generation to affect the precision scores in different kinds of slots. Figure G7 is for the **TEMPLATE ID** slot. The fillers of this slot are arbitrary numbers that uniquely identify the templates for a given message. The scoring program disregards the actual values and finds the best match between the system-generated templates and the answer key templates for a given message based on the degree of match in fillers of other slots in the template. Since there is no such thing as an *incorrect* template ID, only a *spurious* or *missing* template ID, and since missing data plays no role at all in computing precision, the only penalty to precision for the **TEMPLATE ID** slot is due to spurious data generation. In contrast to the **TEMPLATE ID** slot, the **TYPE OF INCIDENT** slot (Figure G9) shows no influence of spurious data on precision at all. This is because the **TYPE OF INCIDENT** slot permits only one filler. The **HUMAN TARGET: ID(S)** slot (Figure G17) can be filled with indefinitely many fillers and thus shows the impact of both incorrect and spurious data on precision.

Four sites submitted results for the optional test runs that were alluded to in the previous section -- BBN Systems and Technologies (BBN-O), New York University (NYU-O1 and NYU-O2), the University of Massachusetts (UMA-O), and the University of Nebraska/University of Southwestern Louisiana (UNL-O1 and UNL-O2). These sites conducted radically different experiments to generate templates more conservatively. The BBN-O experiment largely involved doing a narrower search in the text for the template-filling information; the NYU-O1 and NYU-O2 experiments involved throwing out templates in which certain key slots were either unfilled or were filled with information that indicated an irrelevant incident with good probability; the UMA-O experiment bypassed a case-based reasoning component of the system; and the UNL-O1 and UNL-O2 experiments

involved the usage of different thresholds in their connectionist framework. The experiments resulted in predicted differences in the Matched/Missing scores compared to the basic test. In almost all cases the experiments had the overall effect of lowering recall; in all cases they lowered overgeneration and thereby raised precision. Figure G7 shows the marked difference the experiments made in spurious template generation; Figure G1 shows the much smaller difference they made in overall recall and precision.

CONCLUSIONS

The MUC-3 evaluation established a solid set of performance benchmarks for systems with diverse approaches to text analysis and information extraction. The MUC-3 task was extremely challenging, and the results show what can be done with today's technologies after only a modest domain- and task-specific development effort (on the order of one person-year). On a task this difficult, the systems that cluster at the leading edge were able to generate in the neighborhood of 40-50% of the expected data and to do it with 55-65% accuracy. Breakdowns of performance by slot show that performance was best on identifying the type of incident -- 70-80% recall (completeness) and 80-85% precision (accuracy) were achieved, and precision figures in the 90-100% range were possible with some sacrifice in recall.

All of the MUC-3 system developers are optimistic about the prospects for seeing steady improvements in system performance for the foreseeable future. This feeling is based variously on such evidence as the amount of improvement achieved between the dry-run test and the final test, the slope of improvement recorded on internal tests conducted at intervals during development, and the developers' own awareness of significant components of the system that they had not had time to adapt to the MUC-3 task. The final test results are consistent with the claim that most systems, if not all, may well be still on a steep slope of improvement. However, they also show that performance on recall is not as good as performance on precision, and they lend support to the possibility that this discrepancy will persist. It appears that systems cannot be built today that are capable of obtaining high overall recall, even at the expense of outrageously high overgeneration. Systems can, however, be built that will do a good job at potentially useful subtasks such as identifying terrorist incidents of various kinds.

The results give at least a tentative indication that systems incorporating robust parsing techniques show more long-term promise of high performance than non-parsing systems. However, there are great differences in techniques among the systems in the parsing and non-parsing groups and even among those robust parsing systems that did the best in maximizing recall and precision and minimizing the tradeoff between them. Further variety was evident in the optional test runs conducted by some of the sites. Those runs show promise for the development of systems that can be "tuned" in various ways to generate data more aggressively or more conservatively, yielding tradeoffs between recall and precision that respond to differences in emphasis in real-life applications.

Some conclusions can be drawn regarding the evaluation setup itself that will influence future work. First, the evaluation corpus and task were sufficiently challenging that they can be used again in a future evaluation (with a refined task definition and a new test set). Second, the information extraction task needs modification in order to focus as much as possible on language processing

capabilities separate from information extraction capabilities, and new ideas for designing tests related to specific linguistic phenomena are needed. Finally, more work is needed to ensure that the statistical significance of the results is known, and a serious study of human performance on the task is needed in order to define concrete performance goals for the systems.

ACKNOWLEDGEMENTS

This work was funded by DARPA under ARPA order 6359. The author is indebted to all the evaluation participants, whose collaboration on MUC-3 deserves the highest praise. The author would especially like to thank those individuals who served in special capacities and contributed extra time and energy to ensure the success of the evaluation and the publication of the proceedings, among whom are Laura Blumer Balcom, Nancy Chinchor, Ralph Grishman, Pete Halverson, Lynette Hirschman, Jerry Hobbs, Cheryl Kariya, George Krupka, David Lewis, Lisa Rau, Eric Scott, John Sterling, Charles Wayne, and Carl Weir.

REFERENCES

[1] Grishman, R., and Sterling, J., Preference Semantics for Message Understanding, in *Proceedings of the Speech and Natural Language Workshop*, October, 1989, Morgan Kaufmann, pp. 71-74.

[2] Hirschman, L., Comparing MUCK-II and MUC-3: Assessing the Difficulty of Different Tasks (in this volume).

[3] Chinchor, N., MUC-3 Evaluation Metrics (in this volume).

[4] Chinchor, N., MUC-3 Linguistic Phenomena Test Experiment (in this volume).