OVERVIEW OF THE FOURTH MESSAGE UNDERSTANDING EVALUATION AND CONFERENCE

Beth M. Sundheim

Naval Command, Control, and Ocean Surveillance Center
RDT&E Division (NRaD)[1]
Decision Support and AI Technology Branch
San Diego, CA 92152-5000
sundheim@nosc.mil

INTRODUCTION

The Fourth Message Understanding Conference (MUC-4) is the latest in a series of conferences that concern the evaluation of natural language processing (NLP) systems. These conferences have reported on progress being made both in the development of systems capable of analyzing relatively short English texts and in the definition of a rigorous performance evaluation methodology. MUC-4 was preceded by a period of intensive system development by each of the participating organizations and blind testing using materials prepared by NRaD and SAIC that are described in this paper, other papers in this volume, and the MUC-3 proceedings [1].

The overall objective of the evaluations is to advance our understanding of the merits of current text analysis techniques, as applied to the performance of a realistic information extraction task. As a task, information extraction requires "understanding" of the texts, but it presents a more limited challenge than would a task requiring production of an in-depth representation of the contents of complete texts. The inputs to the analysis/extraction process consist of naturally-occurring newswire texts that were obtained in electronic form. The outputs of the process are a set of templates or semantic frames resembling the contents of a partially formatted database.

MUC-3 and MUC-4 offer benchmarks for the field of NLP in general and for information extraction technology in particular. One of the fundamental ways in which MUC-3 and MUC-4 are distinct from earlier efforts is in their choice of texts: MUC-3 and MUC-4 made use of news articles on the subject of Latin American terrorism, whereas the previous conferences had made use of naval tactical message narratives [2]. The MUC-4 evaluation and conference featured an enhanced evaluation methodology, greater participation, and significantly more conclusive results than those recorded in the MUC-3 proceedings.

Evaluating end-to-end systems in the context of a common task helps in several ways to bridge the gap between research and technology. First, it makes it easier for both technology producers and technology consumers to understand and appreciate the value of the methods that are being explored and applied. It also

[1] formerly the Naval Ocean Systems Center (NOSC).

serves as an example of achievable results and inspires ideas for real-life applications such as statistical analysis or trends analysis of world events, routing/retrieval of texts of personal interest, and feeding of data to expert systems and database management systems. Finally, it encourages the development of large, experimental testbed systems in which to conduct research, and the evaluation results can provide insight into research bottlenecks that are impeding the development of full-scale, usable systems.

This paper presents an overall view of the MUC-4 evaluation and, to a large extent, reflects the content of introductory presentations made at the conference. This paper is also an overview of the conference proceedings, which includes papers contributed by the organizations that participated in the evaluation (Parts II and III) and by individuals who were involved in designing aspects of the evaluation (Part I). The ordering of papers does not necessarily correspond to the order in which the presentations were made during the conference. The proceedings also includes a number of appendices (Part IV) containing materials pertinent to the evaluation.

EVALUATION PARTICIPANTS

Seventeen systems were evaluated for MUC-4, versus 15 for MUC-3. Nineteen organizations participated in the development of the MUC-4 systems, including 12 of the 17 MUC-3 participants. These veteran groups are BBN Systems and Technologies (Cambridge, MA), General Electric (Schenectady, NY), 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), Paramax Systems[2] (Paoli, PA), PRC, Inc. (McLean, VA), SRI International (Menlo Park, CA), the University of Maryland together with ConQuest, Inc.[3] (Baltimore, MD), and the University of Massachusetts (Amherst, MA).

Participation in MUC demands a great commitment of resources over an extended period of time. The actual effort expended for MUC-4 ranged from less than two person-months to over twelve. For the veteran groups, this effort was in addition to the effort spent preparing for MUC-3. Given the commitment required and the limited amount of funding that was available to help support the efforts, it is not surprising that several MUC-3 groups were unable to continue participation in MUC-4[4]. What *is* surprising is that there were seven new MUC-4 participants. These include three organizations currently working under separate DARPA contracts in the area of information extraction, namely Brandeis University (Waltham, MA), Carnegie Mellon University (Pittsburgh, PA), and New Mexico State

[2] formerly Unisys Center for Advanced Information Technology

[3] formerly Synchronetics, Inc.

[4] Those MUC-3 participants that were unable to participate in the MUC-4 evaluation are Advanced Decision Systems (Mountain View, CA), General Telephone and Electronics (Mountain View, CA), Intelligent Text Processing, Inc. (Santa Monica, CA), the University of Nebraska (Lincoln, NE), and the University of Southwest Louisiana (Lafayette, LA).

University (Las Cruces, NM).[5] New participants also included the MITRE Corp. (Bedford, MA), Systems Research and Applications (Arlington, VA), the University of Michigan (Ann Arbor, MI), and the University of Southern California (Los Angeles, CA).

DIFFERENCES BETWEEN THE MUC-3 AND MUC-4 EVALUATIONS

Preparations for MUC-4 were made starting in October, 1991, the call for participation was issued in December, and the system development phase was well underway by February, 1992. A dry run of the evaluation was conducted in late March, final testing was done in late May, and the conference was held in mid-June.[6] The program committee[7] approved an ambitious plan for updating various aspects of the MUC-3 evaluation design for use for MUC-4. Changes to the task definition, corpus, measures of performance, and test protocols were made in order to provide

* greater focus on the issue of spurious data generation;
* isolation of text filtering performance;
* better isolation of language analysis performance;
* assessment of system independence from the training data;
* assessment of system development progress since MUC-3;
* more consistent scoring;
* means to make valid score comparisons among systems.

Greater Focus on the Issue of Spurious Data Generation

The MUC-3 measures of performance implicitly encouraged participants to strive to develop their systems to achieve high recall at the expense of high overgeneration.[8] A few changes were made to the template scoring software to make the generation of spurious data more apparent. One of these changes focuses attention on overgeneration at the slot level (generating more slot values than were expected), while the others focus attention on overgeneration at the template level (generating more templates than were expected).

To address the spurious slot-value issue, an additional method of assessing penalties for missing and spurious data (called the "Matched/Spurious" method) was incorporated, completing the picture provided by the three that had been developed for MUC-3. To address the spurious template issue, a preliminary step in the alignment of response templates with key templates was implemented that requires that minimal "content-based mapping conditions" be met in order for alignment to occur. Response templates that fail to meet these minimal conditions

[5] New Mexico State University teamed with Brandeis University for MUC-4, and Carnegie Mellon University teamed with General Electric.

[6] The conference was hosted by PRC, Inc. at their conference center in McLean, VA.

[7] The MUC-4 program committee included B. Sundheim (NRaD), chair; N. Chinchor (SAIC); R. Grishman (NYU); J. Hobbs (SRI); D. Lewis (U Chicago); L. Rau (GE); C. Weir (Paramax).

[8] Readers unfamiliar with the usage of the terms "recall," "precision," and "overgeneration" as information extraction evaluation metrics should refer to [3].

are scored as spurious; if any unaligned key templates remain, the system gets penalized for missing them. These changes are discussed further in [3].

One way in which the spurious template issue was addressed was to change the way the scoring software does the mapping (or "alignment") of the system-generated "response" templates with the answer "key" templates. A minimum degree of match in the content of the key and response is required before mapping is allowed; if disallowed, the system is penalized for having produced one spurious template (and, under some circumstances, as many spurious slot values as there are values in the response) and for having missed one template (and, under some circumstances, as many slot values as there are values in the key). When multiple template mappings are possible, the scoring program chooses the mapping that is likely to produce the best score. The MUC-3 scoring program used only the latter strategy, the scoring optimization strategy. Thus, no matter how bad the fit in the content of the template, a mapping would be permitted. The MUC-3 method therefore hid the fact that a response template and a key template were representing completely different incidents.

In addition to these changes to the scoring software, the test protocol was modified so that the focus of most of the attention was shifted from the "Matched/Missing" method of scoring, which penalizes at both the template and slot-value level for *missing* information but only at the template level for *spurious* information, to the "All Templates" method, which penalizes at both levels for both types of error. Greater emphasis was also placed on viewing a system's recall and precision as a rectangular "region of performance", whose boundaries are defined by the four methods of assessing penalties. This view of performance reflects the assumption that real-world aplications would vary according to their degree of tolerance of missing data versus spurious data. See test procedure (appendix B) and scatter plots (appendix H).

Isolation of Text Filtering Performance

Overall, approximately 50% of the texts in the MUC-3 and MUC-4 corpora are irrelevant to the information extraction evaluation task. Thus, a significant subtask is to discriminate between relevant and irrelevant texts. MUC-3 scores were computed based on performance at the template level, rather than on the message level, making it difficult to derive a text filtering score. To measure the text filtering capabilities of the MUC-4 systems directly, scores were assigned at the message level and combined using a contingency table. This is discussed further in [3] and [4].

Better Isolation of Language Analysis Performance

Several changes to the template design were made in order to better isolate the systems' capabilities with respect to the kinds of text processing required to meet the differing information extraction requirements (appendix A):

1. Slots in the MUC-3 template that contained composite values were split into two slots. Thus, a MUC-3 slot (**TYPE OF INCIDENT**) filled with the value ATTEMPTED BOMBING became two MUC-4 slots (**INCIDENT: TYPE** and **INCIDENT: STAGE OF EXECUTION**) filled with BOMBING and ATTEMPTED, respectively; similarly, a MUC-3 slot (**HUMAN TARGET: ID**) filled with the value "MARIO

FLORES" ("STUDENT") became two MUC-4 slots (**HUM TGT: NAME** and **HUM TGT: DESCRIPTION**) filled with "MARIO FLORES" and "STUDENT": "MARIO FLORES", respectively.

2. A new string-fill slot (**INCIDENT: INSTRUMENT ID**) was added for identifying the instrument of an attack (e.g., "CAR BOMB"); this slot is paired with the set-fill slot (**INCIDENT: INSTRUMENT TYPE**) that was used for MUC-3 and that now contains a cross-reference to the string-fill slot (e.g., VEHICLE BOMB: "CAR BOMB").

3. New slots (**PHYS TGT: NUMBER** and **HUM TGT: NUMBER**) were added for the number associated with each physical and human target (e.g., 3: "POWER PYLONS" and 4: "ENERGY TOWERS"), supplementing the information in the total number slots (**PHYS TGT: TOTAL NUMBER** and **HUM TGT: TOTAL NUMBER**). The usage of the slots containing total numbers was restricted to cases where the information was not redundant (i.e., to cases where there is more than one such target) and was explicitly mentioned in the text (i.e., cases where no computation by the system is required).

4. The usage of the ATTACK incident type was extended to cover all murder incidents; cases were eliminated in which MURDER templates existed in the training set, either by deletion or by conversion to ATTACK, depending on the circumstances.[9]

5. The slot ordering was changed so that groups of dependent slots appear together, and the scoring software was updated to compute subtotal scores for each group. These groups were termed "pseudo-objects" since they were incorporated as a compromise between retaining the flat template format and replacing it with an object-oriented format. The experimental test designed and conducted by General Electric [5] was an attempt to find out what would have happened if the template format had been overhauled; the pseudo-object computations were essential for that test.

6. The scoring software was updated to include a "STRING FILLS ONLY" row to show how system performance on string-fill slots compares with performance on set-fill slots, for which a "SET FILLS ONLY" row already existed.

Assessment of System Independence from the Training Data

The reuse for MUC-4 of the same domain and fundamentally the same task as used for MUC-3 raised the concern that the "generality" of the systems would come into question. To address these concerns, a controlled generality test was added to the test protocol. The MUC-3 corpus originated from an Foreign Broadcast Information Service (FBIS) archival database containing news articles (from "FBIS

[9] For MUC-3, any incident type other than ATTACK that resulted in the death of one of the human targets was represented in two templates, one of which was a MURDER template. An ATTACK incident that resulted in death to only a subset of the targets was similarly represented in two templates. For MUC-4, these "dependent" MURDER templates were deleted. "Stand-alone" MURDER templates, which were created when the result of an attack was death to *all* targets, were converted to ATTACK templates.

Daily Reports") that had been disseminated as messages [1]. Nearly all those articles carried datelines from 1989 through early 1990; just a few were from 1988.

For the generality test, over 900 different articles of the same varieties as those comprising the MUC-4 corpus were retrieved from a CD-ROM covering August-December 1988, and a sample of 100 was selected as test data and labeled TST4. Sampling factors included the total number of texts for each month in the corpus and, as for the MUC-3 corpus, the total number of texts for each country of interest in the corpus. Thus, the two corpora, including the test sets, report somewhat different incidents and show where the hotbeds of terrorist activity differ from the one time span to the other.[10]

Assessment of System Development Progress since MUC-3

For MUC-3, a study was carried out to measure the complexity of the MUC-3 evaluation task vis-a-vis the previous evaluation, and the scores obtained in the previous evaluation were recomputed using the MUC-3 method of scoring [6]. The evidence was that the MUC-3 task was considerably more complex in most regards and that the MUC-3 scores were about half as good (had twice the shortfall from the upper bound). The conclusion was that the increase in difficulty in the task more than offset the decrease in scores, showing that significant progress had been made.

In the absence of an established, comprehensive methodology, this comparison was necessarily crude since the two evaluations were so different with respect to complexity of the data, corpus dimensions, nature of the task, and scoring of results. In contrast, the differences between MUC-3 and MUC-4 are much less radical, and it was possible to design a controlled comparison between the two. In fact, an attempt was made to neutralize the differences entirely by forward-converting the materials from the MUC-3 final test to the MUC-4 format. Converted materials include the TST2 key and response templates and the cumulative TST2 history file.[11] In addition, the scoring program was configured to disregard those slots in the template that were new for MUC-4 (**INCIDENT: INSTRUMENT ID, PHYS TGT: NUMBER, HUM TGT: NUMBER**) and those that had been incompatibly redefined for MUC-4 (**PHYS TGT: TOTAL NUMBER, HUM TGT: TOTAL NUMBER**).

NRaD rescored TST2 for the MUC-3 veteran sites; the scoring was done noninteractively, using the converted cumulative history file. The MUC-4 test protocol required that all MUC-4 participants do a comparable scoring of TST3, i.e.,

[10] Other differences that existed between the corpora were eliminated. For example, the new corpus was obtained in mixed upper and lower case. TST4 was converted to all upper case in order to be consistent with the original corpus. Also, the new corpus was not stored in the form of messages and, as a consequence, long articles appeared in their entirety rather than being broken up. Any long texts that were selected for inclusion in TST4 were scanned for terrorism key words, and all but a one- to one-and-one-half-page section of text containing one or more of those key words was thrown out.

[11] The history file contains a record of all interactive scoring decisions; the cumulative history file is built up as NRaD scores each system. The scoring program does not query the user if the history covers the case in question. This feature ensures consistency of scoring across systems.

one in which all slots except those mentioned above are scored. NRaD and the MUC-4 participants used the same version of the scoring program (version 3.3).

More Consistent Scoring

The scoring program was updated to further automate the scoring of set-fill slots--the user is now queried only when a set-fill value is cross-referenced to a string-fill value that the scoring program cannot automatically score. It was also updated to score some string fills automatically. The coverage of the interactive scoring guidelines (appendix C) was extended . These updates were meant to ensure greater consistency in template scoring among people and across scoring runs.

The test protocol required that all participants score their own templates. NRaD subsequently rescored the basic test runs for the two new test sets, TST3 and TST4; however, runs such as the one using TST3 to measure progress (described above) were not rescored. In terms of the overall scores for TST3, there was very little difference (0-2% in recall or precision) noted between those that the sites reported and those that were produced when NRaD rescored the outputs. For TST4, the differences ranged from 0-4%.

The actual differences due to subjective scoring are much smaller, however. This is because the rescoring done at NRaD used a slightly updated version of the scoring program (version 3.4a) and a slightly updated version of the answer keys. With respect to the latter, there were more updates made to TST4 than to TST3; hence, the greater range in scoring differences for TST4. As another side note, it is the case that the NRaD overall recall and precision scores are almost always slightly *higher* than those the sites reported; this is probably because NRaD was in a position to interpret the interactive scoring guidelines more liberally than the sites were, while maintaining consistency in subjective decisions across systems.

Means to Make Valid Score Comparisons Among Systems

A well-defined set of evaluation metrics was used for MUC-3, and for the first time, the metrics were implemented as software. This enabled the production of measures of performance at the slot and template levels and measures for subsets of the data (e.g., for only the set-fill slots, for only certain slots in certain templates). With this wealth of data, together with new confidence in the validity of the scores and the maturing state of development of many of the systems under evaluation, there was a growing need for a valid means of making direct cross-system performance comparisons.

For MUC-3, there were no scientific grounds for saying that a system performing at 50% recall and 50% precision was doing "better" than one performing at 30% recall and 70% precision. The only justification for such a claim came from the test protocol, which specified that the run submitted by each site as the system's "required" run be one in which the recall and precision scores were optimized to be as similar as possible. Furthermore, there were no grounds for claiming that a system that got 50% recall and 50% precision was *significantly* better than one that got 48% recall and 48% precision.

Two innovations in the area of scoring were made to address these issues. First, a scientifically sound, single-score measure was incorporated that enabled systems to be ranked. This measure, known as the F-measure, allows different weightings of recall and precision. When they are weighted equally, it does what was only implied by the MUC-3 test protocol, i.e., it would rank a system with 50% recall and 50% precision higher than one 30% recall and 70% precision. Second, a method of doing statistical significance testing was incorporated into the test protocol. This is a computer-intensive method that uses an approximate randomization approach; for MUC-4, it was used for TST3 and TST4 to determine the significance of the overall F-measure scores and All Templates scores. These innovations are discussed further in [3] and [7], respectively.

Shortcomings in the Evaluation

A number of shortcomings in the evaluation remain. In fact, one of the interesting outcomes of MUC-4 was the extent to which the improved system performance brought out the task deficiencies. It is not difficult to define an information extraction task but perhaps even more difficult to make needed improvements without jeopardizing the schedule, placing an undue burden on the evaluation participants, or incurring large costs in terms of updating existing answer key templates and documentation. The compromise reached for MUC-4 was to minimize the changes to the task definition and to focus instead on making improvements to the evaluation metrics and scoring software. Among the remaining shortcomings of the evaluation are the following:

1. The flat template structure created problems as far as meaningfully and consistently expressing inherently recursive kinds of data such as levels of description for perpetrators and human targets. The perpetrator slots allowed for a two-level distinction, with very poor conventions for deciding what to do if the text made more levels of distinction than that, e.g., three levels in "Miguel Vasquez, a member of the Jacobo Carcomo Command of the FMLN". The human target slots had more explicit but still inadequate conventions for entering whatever levels of description were needed to correspond to fillers of other slots, e.g., "five people were injured, including two security guards". Another consequence of the flat template structure was the requirement to encode explicit cross-references, greatly complicating the scoring algorithms.

2. The definition of a "relevant terrorist incident" was inadequate in several respects. The distinction between a terrorist incident and other kinds of violent events -- guerrilla actions, shootings of drug traffickers, etc. -- is a difficult one to express in comprehensive, hard and fast rules. It was also difficult to express the relevance criteria of "specificity" and "recency" in a way that could be consistently applied. The intent was to not do extraction unless some specific information was present that a database user would find useful; for example, extraction would not be done if no particular incident was being referred to ("terrorist bombings have been taking place with increasing frequency", "over 100 bombings have taken place in the last two weeks"). If an incident was reported as having taken place more than two months prior to the date of the article, no extraction was to be done unless the article gave "new" template-filling information, e.g., when a new suspect was being brought forth. However, without prior knowledge of the actual incident, it was sometimes difficult to tell whether the information that was being reported was new or not. These problems of determining relevance were partly due to the task definition and partly due to the inherent vagueness of the texts.

3. There were small gaps in the template fill rules. For example, the rules concerning stories that give contradictory evidence about some of the facts were inadequate. A more frequent problem was that the set-fill lists for physical and human target types were sparse and sometimes vaguely defined, and some of these problems had consequences for determining relevance at the template level. For example, if a text describes the target of an incident only as a "naval attache", the incident is relevant if the target is classified as DIPLOMAT but irrelevant if the target is classified as ACTIVE MILITARY.

4. In terms of the scoring, there were several relatively minor but troublesome problems. A bug in the scoring program was discovered just prior to final testing, and a change was made to the scoring program and the interactive scoring guidelines just prior to final testing that had to be retracted when NRaD rescored TST3 and TST4. The largest number of problems were those that involved making subjective judgments during interactive scoring. For example, string fills that closely resembled the ones in the key but originated from remote places in the texts had to be examined in context to determine whether they were "fortuitously correct" (as, perhaps, in the case of "urban guerrillas" as a substitute for "urban terrorists") or "infortuitously incorrect" (as in the case of "11 peasants" as a substitute for "3 peasants"). Making principled decisions about awarding partial credit was also difficult when the cases weren't specifically covered by the interactive scoring guidelines.

5. The change in template mapping strategy described earlier as an improvement made for MUC-4 had one consequence that was at least potentially problematic. The problem is due to the inflexibility of one of the mapping conditions, namely the requirement that there be at least a partial match on the filler for **INCIDENT: TYPE**. A partial match existed when the response was ATTACK, and the key was any other value; this scoring is based on ATTACK being a supercategory of the other set-fill options. In the reverse case, however, the response is scored incorrect, thereby disallowing the mapping and, as described earlier, resulting in penalties for having generated a spurious template and for having missed a template. The disallowance of a mapping simply on the basis of an incorrect incident type is probably too extreme. (In practice, however, it appears to have rarely had significant adverse consequences; see UMass paper in Part II as one example of it having apparently significantly affected their TST4 performance.)

At a higher level, there are shortcomings that are due to the choice of task. Information extraction has served as an excellent vehicle for elucidating the application potential of current technology; however, its utility as a vehicle for focusing attention on solving the "hard" problems of NLP is not as great. Many insights have been gained into the nature of NLP by experience in developing the large-scale systems required to participate in the evaluation. Nevertheless, so much effort is involved simply to make it through the evaluation that it takes a disciplined effort to resist implementing quick solutions to all the major issues involved, whether they are well understood problems or not.

The attempts that have been made to use the information extraction task to reveal language analysis capabilities specifically have so far met with limited success. One of these examined the results of information extraction at the local level of processing (apposition handling), and the other looked at the global level of processing (discourse handling). The former was carried out for MUC-3 [8] and the latter for MUC-4 [9]. The major conclusions of the apposition test were that the test was isolating the phenomenon to some extent and that the systems as a group were doing better on the cases that had been hypothesized as easier than on those that had been hypothesized as more difficult. However, it also appears that performance on the apposition test may have reflected the systems' slot-filling capabilities at least as much as their apposition analysis capabilities. Apposition was chosen as the subject of the test partly because of the relatively high frequency of occurrence of the phenomen; however, a substantial portion of the cases introduced confounding factors and had to be thrown out. The major conclusion of the discourse processing test was that the texts that were expected to be "easy" were not and that there was something about the composition of the small test samples that were used that was confounding the results. Although there seems to be no theoretical impediment to conducting successful fine-grained task-oriented tests, these two efforts seem to show that such tests cannot be designed as *adjuncts* but rather require independent specification in order to ensure adequate test samples and an appropriately designed information extraction task.

DISCUSSION OF TEST SETS AND TEST RESULTS

Appendix B describes the performance objectives of the evaluation, the components of the test, how the sites were to conduct the tests and score the outputs, and what files the sites were to submit to NRaD after finishing the test procedure. Appendix G contains summary score reports for the component tests, and appendix H displays some of those results in the form of scatter plots. The discussion below concerns the results for the basic test components, namely TST3, TST4, and the TST2/TST3 "progress" test. The "adjunct" tests that are mentioned in appendix B are reported on in [5] and [9].

TST2/TST3 "Progress" Test

The progress test made a controlled comparison between MUC-3 and MUC-4 performance. The data points for MUC-3 were obtained using the templates that the veteran participants' systems generated on the MUC-3 final test on TST2. The data points for MUC-4 were obtained for *all* MUC-4 sites; they were obtained using the templates generated on TST3. As described earlier, the TST2 test materials were forward-converted to the MUC-4 format, and scoring included only those template slots whose MUC-3 and MUC-4 definitions were consistent.[12]

[12] The TST3 progress scores are generally slightly better (up to 2%) than TST3 "base" scores; this difference is the result of having excluded the number slots and the instrument ID slot from the scoring on the progress test.

The TST2 progress scores are generally substantially worse (at least 5% lower recall or precision) than the MUC-3 TST2 "base" scores reported on in [3]. The changes (primarily decreases) are due to such factors as the following:

1. The manual clean-up of the automatic forward conversion of the templates is subject to a small degree of error. The elimination of some MURDER templates via conversion to ATTACK templates could result in an underestimation of performance; the splitting of the human target ID information into two slots could result in an overestimation of performance.

2. Since scoring of the TST2 templates for the progress test was done in batch, without any manual template remappings, performance may be slightly underestimated for the few sites

Following are some of the hypotheses that were to be tested concerning the performance of the MUC-3 veteran systems:

1. Most MUC-3 veteran systems would improve on at least one measure.

2. Systems that were at the leading edge of performance for MUC-3 might not be able to attain higher scores on one measure without sacrificing performance on another.

3. The limitations of some approaches might emerge.

4. The need for progress in certain research areas might become salient.

5. The fairest (and most generous) view of progress would come from the Matched/Missing row, which was the focus of the MUC-3 test, rather than from the more stringent, All Templates row, on which the MUC-4 TST3 and TST4 tests focused.

A comparison of the tables in section 4 of appendix G shows improvement on one of the three primary measures (recall, precision, overgeneration) by all 11 systems, given the Matched/Missing method, and by 10 of the 11 systems, given the All Templates method. Improvements on all three measures were achieved by three systems on Matched/Missing (GE, LSI, NYU), including two of the leading MUC-3 performers (GE, NYU), and by seven systems on All Templates (GE, LSI, NYU, PRC, SRI, UMBC-ConQuest, UMass), including several of the leading performers (GE, NYU, SRI, UMass). Tradeoffs resulting in improved recall at the expense of lower precision are evident in the results for three systems on Matched/Missing (Hughes, Paramax, UMass) and for one system on All Templates (Paramax). Tradeoffs leading to improved precision at the expense of lower recall can be seen in the results for two systems on Matched/Missing (BBN, MDC) and one system on All Templates (BBN).

	M/M&AT REC	M/M PRE	AT PRE	M/M OVG	AT OVG
MAX CHG FOR WORSE	-7	-10	-19	+14	+26
MAX CHG FOR BETTER	+22	+10	+34	-8	-42
AVERAGE CHG	+8	+2	+11	+2	-12

Table 1. Differences Between TST2 and TST3 Scores on Progress Test

The differences in scores between the two test sets are summarized in Table 1. The differences are calculated as the TST3 progress score minus the TST2 progress score. The first row shows the worst degradation among the 11 systems, the second row shows the most improvement, and the third row shows the average change.

that made substantial use of this facility; however, the need for this facility has declined as the template alignment capabilities of the scoring program have improved.

[3] The elimination of some MURDER templates via deletion eliminated one source of inflation of scores.

[4] The scoring program now uses more stringent criteria when aligning templates; the impact is generally a higher missing template count, which lowers recall, and a higher spurious template and slot-filler count, which lowers precision.

Note that there is only one column for recall, which is unaffected by the choice of Matched/Missing (M/M) versus All Templates (AT).

Recall improved by an average of eight percentage points. On average, there was very little change in precision and overgeneration on Matched/Missing, but All Templates shows dramatic improvement on both measures. It is interesting that the progress is more evident on All Templates than on Matched/Missing: for nine of the eleven systems, the All Templates precision and overgeneration scores show a larger improvement from MUC-3 to MUC-4 than do the Matched/Missing scores. It appears that the new focus on the All Templates row caused developers to devote a great deal of attention to reducing overgeneration (thereby increasing precision), and that they succeeded. Furthermore, of these nine systems, eight showed improved recall as well.

The F-measures provide assistance in interpreting the results of this test, especially for those systems that exhibited a recall-precision tradeoff. The F-measure scores show whether or not the tradeoff paid off in terms of overall performance. Figure 1 shows the MUC-3 veteran systems' All Templates F-measure scores (with recall and precision equally weighted) from the tables in appendix G, section 3.

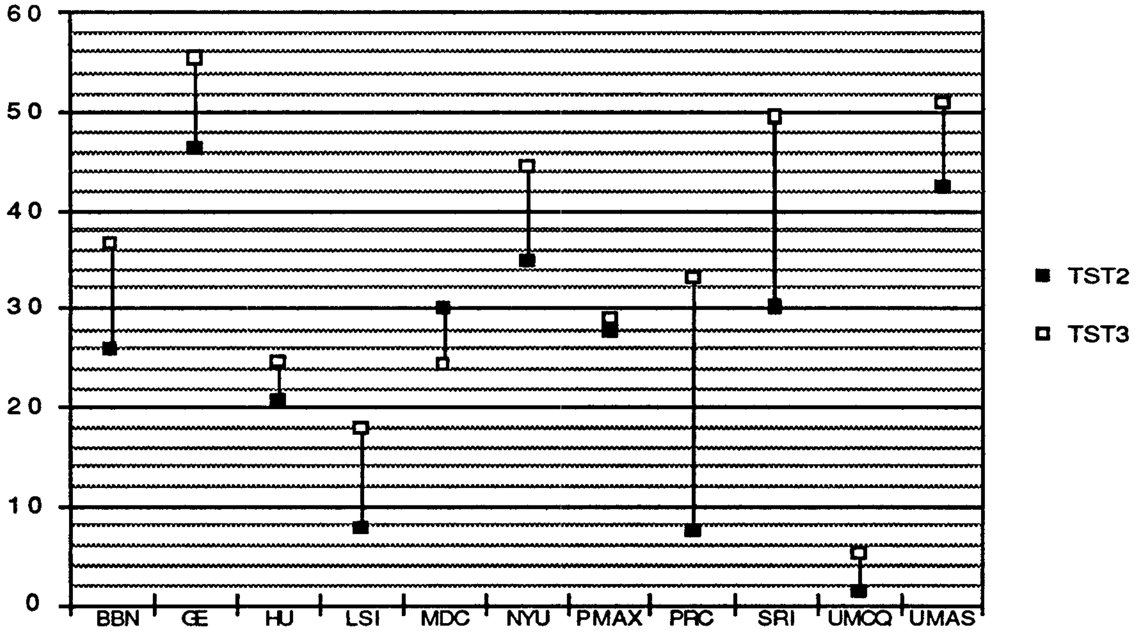

Figure 1. All Templates F-Measure (P&R) for Progress Test on TST2 and TST3

Figure 1 shows that a "typical" increase in F-measure performance is around 10 points (BBN, GE, LSI, NYU, UMass), and two systems (PRC and SRI) show a much greater performance improvement than that. The SRI results are especially remarkable because of the radical differences between their MUC-3 and MUC-4 systems. The BBN results show that the tradeoff in performance they made for MUC-4 clearly paid off in terms of overall progress.

Two of the systems exhibiting only a slight performance increase (Hughes, Paramax) do little or no linguistically-based processing; by their developers' own admission, the systems are incapable of much higher performance unless they are augmented by other types of processing. The remaining two systems, MDESC and UMBC-ConQuest, were overhauled for MUC-4. In the case MDESC, this overhaul resulted in lower overall performance than what was achieved for MUC-3; in the case of UMBC-ConQuest, it resulted in a modest increase but still very low overall performance. It should be noted that the level of effort that could be afforded by each of these four sites was minimal and that this undoubtedly was a significant limiting factor.

Systems representing organizations that are not veterans of the MUC-3 evaluation are not included in the above discussion. They were tested on the TST3 portion of the progress test. Their scores are included in appendix G, section 4.

In summary, the progress test showed that higher levels of performance by nearly all systems were achieved despite the relative difficulty of TST3. Progress was more evident when the All Templates scores are considered; this is due to the success of most systems in controlling overgeneration. Most systems did not give evidence of a recall-precision tradeoff, which means that there is still a variety of techniques that exhibit potential for attaining even higher levels of performance in the future. The few systems that exhibited a tradeoff clearly benefited from it in terms of overall performance. However, minimal improvement was shown by systems that do not use linguistically-based processing, and minimal progress or even a degradation in performance were the result in a couple cases where systems were radically changed for MUC-4.

TST3 and TST4 Tests

This section describes the "base" MUC-4 tests, which used the TST3 and TST4 test sets. As distinct from the progress test discussed in the previous section, the base tests scored the entire template rather than selected slots. Thus, the TST3 scores for the two tests can be different, but in reality differences turned out not to be universal. Where differences do exist, they are fairly small -- the overall recall, precision, and overgeneration base scores are at most three points lower than the progress scores.

As described earlier in this paper, TST3 consists of a sample of 100 previously unseen texts from the corpus of FBIS texts that had been obtained prior to MUC-3. The sampling method ensures that the test set contains the same percentage of texts by country as the corpus as a whole; aside from enforcing that constraint, sampling is done blindly. The TST4 test set consists of a sample of 100 texts from the new corpus of FBIS texts that was obtained via CD-ROM specifically for MUC-4.

The density of relevant information in TST3 is relatively high, making it in some ways a more difficult test set than others. The density of relevant information in TST4 is much more similar to TST1 and the training set than it is to TST2 and TST3, making it in some respects a relatively simple test. Some of the differences between TST3 and TST4 are summarized below.

1. Approximately two-thirds of the texts in TST3 (65 out of 100) fall in the "definitely relevant" category, versus approximately one-half in TST4 (48 out 100).

2. Almost one-half the texts in TST3 (30 out of 65) require the generation of more than one template, versus almost one-third in TST4 (15 out of 48).

3. Many templates include a greater density of information than usual, especially in such slots as **HUM TGT: DESCRIPTION**.

In reality, TST3 is just a bit more difficult by each of these criteria than TST2, which was used for final MUC-3 testing.[13] However, TST2 was itself more difficult than TST1 and the training set.[14]

As mentioned earlier, the purpose of introducing the TST4 test was to learn to what extent system performance is independent of the training data. The variable introduced by TST4 was the time span covered by the texts. The change in time span meant that a somewhat different set of incidents would be reported -- no incidents occuring later than 31 December, 1988 would be reported in TST4, whereas incidents up through early 1990 would be reported in TST3. It also meant that the incidents would reflect a different world situation, resulting in a different distribution of articles among the countries of interest. The major differences in this respect were in the number of articles about El Salvador (down from 40% in TST3 to 25% in TST4), Chile (up from 5% to 18%), and Peru (up from 6% to 19%).

The hypotheses to be tested were that systems would not perform as well on TST4 as on TST3 and that systems that rely more heavily on corpus-based statistics would suffer a greater hit in performance than other systems. The results, however, are mixed with respect to the first hypothesis and apparently negative with respect to the second. Table 2 presents a summary of the All Templates scores for the base runs on TST3 and TST4 (appendix G, sections 1 and 2), including the floating point F-measure with recall and precision equally weighted (appendix G, section 5).

	REC	PRE	OVG	F-MEAS (R&P)
TST3 BEST	58	55	26	56.01
TST4 BEST	62	53	34	57.05
TST3 WORST	2	8	90	4.47
TST4 WORST	3	10	87	5.79
TST3 AVG	31	34	55	31.35
TST4 AVG	35	33	57	32.26

Table 2. Summary of TST3 and TST4 All Templates Scores (Base Test)

The TST3 scores are quite similar to the TST4 scores, despite the differences noted between the test sets. Naturally, however, the degree of similarity varies

[13] Several participants did not use TST2 to train on for MUC-4; instead, they reserved it for use as blind test data for internal tests. When reported in the papers in Part II (e.g., by SRA and SRI), the results seem to confirm the degree of similarity between TST2 and TST3, in the sense that the systems did just slightly worse on TST3 than on the last internal test run on TST2.

[14] A table of some summary statistics concerning all four test sets and the training set is included in the BBN paper in Part II.

from one system to the next. Inspection of the individual systems' scores shows that only two systems (LSI, UMich) had both lower recall and lower precision on TST4 than on TST3, and the degradation in recall for LSI is only one percentage point. For two other systems (PRC, SRI) recall was the same on both test sets, while precision was lower on TST4. Eleven systems showed higher recall and lower precision on TST4. Two systems (USC, MITRE) scored higher recall *and* higher precision on TST4. Where there was a difference in recall or precision, the degree of difference is as great as 11 recall points and 13 precision points (cf appendix H, figure H7).

The F-measure value (with recall equally weighted with precision) is higher on TST4 than on TST3 for 10 of the 17 systems (BBN, GE, GE-CMU, Hughes, MDESC, MITRE, NYU, SRA, UMBC-ConQuest, USC), is less than two points lower on TST4 for four others (NMSU-Brandeis, Paramax, PRC, SRI), and is more than two points lower on TST4 for only three systems (LSI, UMass, UMich). The absolute rankings (without considering whether the differences are statistically significant) show six systems ranked the same on both test sets, ten changing rank by just one position, and one changing rank by two positions. Thus, in a very real sense, the differences in performance from a cross-system perspective are minimal, and it can be concluded that the two test sets are giving consistent results.

The overall performance of more than half the systems was better on TST4 than on TST3, as determined by the F-measures. The relative straightforwardness of the TST4 test set may have washed out or even reversed the predicted behavior with respect to recall.[15] The expected negative effect of using a corpus spanning a different period of time was not seen; it would be necessary to place more controls on the information density characteristics of the test sets in order to isolate such a factor.

BBN, GE, NYU, SRI, and UMass submitted the results of optional tests conducted using TST3 or both TST3 and TST4. The optional tests explored ways of controlling system behavior to produce recall-precision tradeoffs that were predicted to be suboptimal overall (compared to the base run) but distinctly better on one measure or the other. These tests varied greatly in their design and in the performance impact; further information is available in appendices G (section 3) and H (figures H3, H4, H9) and in the papers in Part II.

A few general comments can be made on the basis of the scatter plots in appendix H concerning the overall performance of the systems. Figures H1 and H2 show that higher recall is usually correlated with higher precision, just as the MUC-3 results showed. Therefore, once again there is no reason not to be optimistic about seeing continued improvement on both measures in the future. Figures H5 and H6 plot overall recall versus overgeneration; they show that, to a large extent, the overall precision scores seen in H1 and H2 are accounted for by the overgeneration factor. This shows that overgeneration is still a serious problem, although MUC-4 clearly demonstrated that a great deal of progress had been made in this area. Clearly also, the problem of missing information is still serious, as witnessed by the fact that recall is still only moderate.

[15] With respect to precision, it should be noted that the two systems that showed better recall and precision on TST4 than on TST3 (USC and MITRE) are less mature than most, which may make their performance less predictable.

The question of how to assess the state of the art has to be addressed in part by comparison to human capabilities, since the real-life challenge is still for systems to try to match the performance of well-trained people. Although the human performance limits have not been scientifically determined, they are now estimated to be in the neighborhood of 75% recall and 85% precision, assuming the All Templates scoring method and a representative test set. These figures may seem low; however, the experience of generating the key templates for these evaluations suggests that they are not. Human factors play a role in estimating this limit; however, the major factors are the task deficiencies and the inherent ambiguity and vagueness of the texts. These performance goals mean, therefore, that the leading systems are falling perhaps 15% short of the recall target and 30% short of the precision target.

Figures H10-H12 plot the "regions of performance" of the systems as defined by the overall Matched/Missing, Matched/Spurious, Matched Only, and All Templates recall and precision scores. There are some interesting differences in the shape as well as the size of those regions. For the systems displaying the smallest regions of performance (H10), the shape is rather square, or it is elongated more horizontally than vertically. In contrast, the regions in H11 and to an even greater extent the regions in H12 are distinctly rectangular and elongated vertically. These shapes are evidence that the systems in H10 are least affected by overgeneration; those in H12 are most affected. There is some comparative proof that the MUC-3 veterans were bringing overgeneration under control in the fact that H12 includes only one veteran system

Figures H15-H18 show that, as anticipated, system performance on slots requiring string fills would be worse than on those requiring set fills. The differences would probably be more striking if it were not for the fact that the scoring of eight of the eleven set-fill slots is confounded by the cross-references attached to them. (In contrast, just one of the six string-fill slots has a cross-reference requirement.) Whether for this reason or not, it does not appear that the distinction in slot type serves as a discriminator among systems, since there are no dramatic differences in the relative position of the systems in the contrasting graphs across both test sets.

GENERAL OBSERVATIONS

There are many ways in which MUC-4 has surpassed MUC-3 in bringing various aspects of the evaluation into focus, including the deficiencies remaining in the task that were described earlier. The challenge posed by the task appears less imposing now -- it is now the rule rather than the exception to find systems capable of exploiting the large training corpus of texts and templates for the purposes of knowledge acquisition, automatic training, and internal testing. The interaction between systems engineering concerns and theoretical concerns is receiving increasing attention. In particular, scalability and robustness issues must be addressed in order to take full advantage of the corpus for training purposes and to perform as well as possible on new test data.

Whereas the challenge posed by the task has come to be accepted more or less as a matter of course, the burden of preparing for the evaluation is increasingly felt. Beneficial effects of the task challenge and evaluation burden are, among other things, that the algorithms for dealing with large amounts of unrestricted text have become more robust, the development cycle has gotten shorter, and the amount of automated knowledge acquisition has increased. On the down side, the evaluation burden is still such that quantifiable progress is slow; there is still a strong sentiment that time is the primary limiting factor, not technology, and that therefore level of effort is one of the most significant factors in predicting performance.

However, even though one impediment to improved performance is the amount of time that can be invested in just doing a lot of hard work, including a great deal of knowledge engineering and system engineering, it is even more apparent from MUC-4 than from MUC-3 that there are certain prevalent "hard problems" posed by the task that require serious study. One thing that has been noted is how small problems in early stages of processing can have large negative effects on the ability of later stages to do their job. MUC-3 (and earlier evaluations) pressed the point of reducing the fragility of sentence-level processing, and the sentence analyzers were developed to produce output even when they didn't have full coverage. MUC-4 has refocused attention on the sentence and the importance of doing more complete linguistic analysis at that level.

Another thing that has become nearly universal experience is the inadequacy of current approaches to determine when and how to combine information from multiple sentences into a single, coherent representation. Although the approaches are limited in effectiveness by the quality of the sentence-level interpretation, they are also inherently limited in their ability to incorporate information from sentences that lack domain-specific "key words", to incorporate information from anaphors (especially from definite noun phrases), and to deal with interruptions in the discourse. Currently these discourse phenomena are generally dealt with in terms of template "splitting" and "merging" based on the compatibility of data in the output representation rather than by tracking discourse as part of the analysis process. Some of these issues are apparent in the participants' discussion in Part III of the "system walkthrough" example (appendix F).

The techniques that were used to improve performance above MUC-3 levels still vary greatly, but the emphasis on hybrid systems combining linguistic and nonlinguistic processing has increased, and the limitations of the purely nonlinguistic approaches are very evident. As the viability of information extraction as a useful application of NLP has increased, the idea of building systems specifically for that purpose has emerged, and there is beginning to be a division between those who would insist that the most successful systems will be the most generic ones with respect to application task and domain and those that believe that the most successful systems will take advantage of whatever reductions in level of sophistication are permitted by the task of information extraction. At the bottom is the question of what it will take to get from the current limit of about 60% recall and 55% precision to the estimated upper limits of human performance. Also at issue is the issue of portability in terms of system architecture and portability in terms of cost. Will it cost less to port a large, complicated system that has separate domain-specific modules to a new domain and/or task, or will it cost less to port a smaller, simpler system to a new domain and to build a new system for a new task?

CONCLUSIONS

New performance standards were set on the MUC-3 and MUC-4 information extraction task. Despite increased task difficulty and scoring stringency for MUC-4, the results of a MUC-4 test to measure progress since MUC-3 show substantially higher overall performance for most systems (at least 10 points higher on the F-measure). It has now proven possible to achieve overall scores above 60% recall and 55% precision and an F-measure exceeding 55. The new challenge to control overgeneration was successfully met, although overgeneration is still high enough that it exerts a major negative impact on precision.

The results of a test to measure the generality of the MUC-4 algorithms show that they were not overly tuned to the training set. The usage of a test set from a corpus spanning a different period of time than that of the original corpus was expected to have a negative effect on performance, but this effect was not seen. It would be necessary to place controls on the information density characteristics of the test sets in order to isolate the time factor.

Upper limits on human performance of the task are estimated to be 75% recall and 85% precision, primarily due to deficiencies in the task definition and expressiveness of the formalism and to the inherent ambiguity and vagueness of the texts. System performance falls short of these levels by at least 15 recall points and 30 precision points. However, some MUC-4 systems attained high enough performance that task deficiencies account for a significant portion of the penalties incurred by the scoring.

Clearly, the performance envelope could have been pushed out even farther if the participants had had the opportunity to work on the systems steadily for the entire year. The level of effort is reflected to some extent in the scores, and time was again a limiting factor. The differences in sophistication among the systems may be great, but these differences may not be so great in terms of the scores. However, it could well be that there is a great qualitative difference between an F-measure score of 45 and one of 55. Since the task deficiencies are being raised as a limiting factor and certain theoretical issues such as those involving sentence- and discourse-level analysis are becoming limiting factors as well, it may be possible to conclude that the ceiling on performance is much more perceptible than it was after MUC-3 and that major steps forward in the state of the art may not be easy to obtain.

Error analyses point toward the critical need for research in areas such as discourse reference resolution and inferencing. For example, the inability to reliably determine whether a description found in one part of the text refers or does not refer to something previously described inhibits both recall and precision because it could result in either missed information or spurious information; the inability to pick up subtle cues to relevant information places a limitation on recall because it results in missed information. The ability to take advantage of sophisticated approaches to discourse that have already received computational treatment is limited by their dependence on error-free outputs from earlier stages of processing. There is a need for renewed attention to robust processing at the sentence level.

It is time to move on to a different information extraction task and domain in order to make further progress in the evaluation methodology and to ensure that the challenge to handle unrestricted text remains high. MUC-4 has clarified many of the issues pertaining to the definition of a performance evaluation using an information extraction task; at some point, it will be worthwhile to try to design a more comprehensive performance test of NLP capabilities than what the information extraction task covers.

ACKNOWLEDGEMENT

All the evaluation participants are to be congratulated for their efforts in support of MUC-4. I would like to thank Nancy Chinchor, Ralph Grishman, Jerry Hobbs, David Lewis, Lisa Rau, and Carl Weir for their commitment as program committee members to improving the evaluation methodology and for their attention to the deluge of email communications. Thanks also to PRC, Inc., for hosting the conference and to Richard Tong and Lynette Hirschman in their roles as "outside evaluation experts". The NRaD work was supported by DARPA/SISTO under ARPA order 6359.

REFERENCES

[1] *Proceedings of the Third Message Understanding Conference (MUC-3)*, May, 1991, Morgan Kaufmann.

[2] Sundheim, B., Plans for a Task-Oriented Evaluation of Natural Language Understanding Systems, in *Proceedings of the Speech and Natural Language Workshop*, February, 1989, Morgan Kaufmann, pp. 197-202.

[3] Chinchor, N., MUC-4 Evaluation Metrics (in this volume).

[4] Lewis, D. and Tong, R., Text Filtering in MUC-3 and MUC-4 (in this volume).

[5] Krupka, G. and Rau, L., GE Adjunct Test Report: Object-Oriented Design and Scoring for MUC-4 (in this volume).

[6] Hirschman, L., Comparing MUCK-II and MUC-3: Assessing the Difficulty of Different Tasks, in *Proceedings of the Third Message Understanding Conference (MUC-3)*, May, 1991, Morgan Kaufmann, pp. 25-30.

[7] Chinchor, N., Statistical Significance of MUC-4 Results (in this volume).

[8] Chinchor, N., MUC-3 Linguistic Phenomena Test Experiment, in *Proceedings of the Third Message Understanding Conference (MUC-3)*, May, 1991, Morgan Kaufmann, pp. 31-45.

[9] Hirschman, L., An Adjunct Test for Discourse Processing in MUC-4 (in this volume).

[10] Sundheim, B., Overview of the Third Message Understanding Evaluation and Conference, in *Proceedings of the Third Message Understanding Conference (MUC-3)*, May, 1991, Morgan Kaufmann, pp. 3-16.