SURVEY OF THE MESSAGE UNDERSTANDING CONFERENCES

Beth M. Sundheim

Naval Command, Control & Ocean Surveillance Ctr.
RDT&E Division (NRaD), Code 444
San Diego, CA 92152-7420

Nancy A. Chinchor

Science Applications International Corporation
10260 Campus Point Drive, M/S A2-F
San Diego, CA 92121

ABSTRACT

In this paper, the Message Understanding Conferences are reviewed, and the natural language system evaluation that is underway in preparation for the next conference is described. The role of the conferences in the evaluation of information extraction systems is assessed in terms of the purposes of three broad classes of evaluation: *progress*, *adequacy*, and *diagnostic*. The conferences have measured system performance primarily to assess progress and the state of the art, but they have also been influenced by the concerns associated with assessing adequacy and providing diagnostics. Challenges for the future of similar evaluations are also discussed.

1. INTRODUCTION

Much has happened since the last time a paper appeared in the ARPA workshop proceedings about the Message Understanding Conferences [11]. The evaluation methodology has been changing steadily, and more demanding information extraction tasks have been defined. In response to the challenges of the evaluation task and metrics, researchers have developed robust and efficient methods for working with large corpora and have confronted prevalent text analysis issues that have so far constrained performance.

These challenges have also resulted in a critical rethinking of assumptions concerning the ideal system to submit for evaluation. Is it a "generic" natural language system with in-depth analysis capabilities and a well-defined internal representation language designed to accomodate the translation of various kinds of textual input into various kinds of output? Or is it one that uses only shallow processing techniques and does not presume to be suitable for language processing tasks other than information extraction?

2. REVIEW OF PAST MUCs

The first Message Understanding Conference (MUC) was held in 1987, used ten narrative paragraphs from naval messages as a training corpus and two others as test data, and had no defined evaluation task or metrics. Researchers from six organizations ran their systems on the test data during the conference, then demonstrated and explained how the systems analyzed the texts. Two years later, the second MUC was held [10]. It made use of a training corpus of 105 naval message narratives of four different types, a dry-run test set of 20 narratives, and a final test set of five. An information extraction task was defined that consisted of identifying ten different pieces of information and representing them as slot fillers in a *template* resembling a semantic frame. This task emulates an information management application requiring the culling of facts from a large body of free text as a means to generate updates to a formatted database.

A rudimentary set of scoring standards was developed, and the templates produced by the eight systems (including four of the six systems represented at the 1987 evaluation) were scored by hand by comparison with a hand-generated answer key. The nature of the corpus used for the second MUC was difficult enough that grammar coverage and parsing efficiency were serious issues. The domain was complex enough that the knowledge engineering job was greatly facilitated by the availability of documentation presenting much of the essential, declarative domain knowledge in a structured format.

After another two-year interval, MUC-3 was held in May, 1991, followed by MUC-4 in June, 1992. There are published proceedings for the third and fourth conferences [8, 9], including descriptions and test results of the participating systems (15 for MUC-3, 17 for MUC-4). A new corpus of 1,400 texts on the subject of Latin American terrorism was used that includes 16 text types (transcribed speeches, newspaper articles, editorial reports, etc.). The template developed for MUC-3 contained slots for 17 pieces of information; the number of information-bearing slots increased to 22 for MUC-4. The scoring metrics were refined and implemented for MUC-3 and MUC-4 in a semiautomated scoring program.

For MUC-3, a study was carried out to measure the complexity of the MUC-3 terrorism task vis-a-vis the naval task, and the scores obtained in the 1989 evaluation were recomputed using the MUC-3 method of scoring [5]. Although these scores were lower, the conclusion was that significant progress had been made, because the increase in difficulty in the task more than offset the decrease in scores.

It was possible to conduct a more refined study of the progress from MUC-3 to MUC-4 [12] that showed that higher levels of performance by nearly all veteran systems were achieved despite the relative difficulty of the MUC-4 test set that was used in the comparison and despite increased strictness of the scoring with respect to spurious data generation. The results of MUC-4 show that higher recall is usually correlated with higher precision[1], which is consistent with the results of previous evaluations and suggests that there is still a variety of techniques with potential for attaining even higher levels of performance in the future. In absolute terms, however, recall and precision scores were still only moderate.

According to an analysis of the effectiveness of techniques used by MUC-3 systems [4], pattern-matching techniques (with hand-crafted or automatically acquired patterns) and probabilistic text categorization techniques proved successful only when combined with linguistic techniques. The use of robust processing including robust parsing was shown to correlate with the success of the system. In a comparison of MUC-3 and MUC-4 systems, minimal improvement from MUC-3 to MUC-4 was demonstrated by the two systems that did not use linguistically-based processing [12]. Several linguistically-based MUC-3 systems improved considerably via extensions made for MUC-4, as did one MUC-3 system that was converted from a generic text understanding system to an information extraction system that maintains its basis in linguistics but is streamlined for speed and geared specifically to the demands of information extraction. However, other systems which underwent a complete overhaul for MUC-4 showed only slight progress or even a degradation in performance.

Error analyses point to the critical need for further 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 the system either missing information or generating spurious information; the inability to pick up subtle relevance indications (e.g., that persons described as being "in" a place that was attacked could be targets of the attack) and not-so-subtle ones (e.g., that a vehicle whose roof collapsed as a result of a bomb explosion was damaged by the explosion) 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 a dependence on error-free outputs from earlier stages of processing. Thus, there is a need for renewed attention to robust processing at the sentence level.

[1] Recall is the ratio of correctly generated fills to the total number of expected fills; precision is the ratio of the correctly generated fills to the total number of generated fills. Thus, shortfalls in recall reflect the amount of missing fills as well as incorrect fills, and shortfalls in precision reflect the amount of spurious fills as well as incorrect fills. See [2] for detailed information on the formulation of these and other metrics, which are under review for MUC-5.

3. MUC-5

We are in another one-year cycle this year, with MUC-5 scheduled for August, 1993. Over 20 organizations are currently planning to participate in the evaluation. Among the expected participants are the organizations already working on the Tipster Text extraction program, other MUC-4 veteran organizations, and six additional participants, four of whom are from outside the United States.

The final evaluation of the Tipster contractors' systems will be the MUC-5 evaluation. There are four tasks, each with its own corpus: joint ventures in English and in Japanese and microelectronics in English and in Japanese. The Tipster-sponsored organizations will be evaluated on all tasks that they are contracted to work on; other MUC-5 participants are allowed to work on both languages if they want to but have been required to choose between the two domains to keep them from spreading their efforts too thin.

The joint ventures task (in both languages) appears to pose significantly greater challenges than the microelectronics task, largely because the joint ventures articles are less technical and more varied in style, are generally longer, and often discuss more than one joint venture. The template includes over 40 content-bearing slots identifying and interrelating various facts about the joint venture and the entities involved. The microelectronics template has fewer slots; it covers features of microchip fabrication processes and the organizations mentioned in association with those processes.

4. ROLES IN EVALUATION

Three broad types or purposes of evaluation have been identified and described by H. Thompson and M. King[2]: *progress evaluation*, *adequacy evaluation*, and *diagnostic evaluation*. The MUC evaluations have been primarily examples of progress evaluation, which is defined as "assessing the actual state of a system with respect to some desired state of the same system, as when progress of a project towards some goal is assessed." However,

[2] These were outlined by Henry Thompson (University of Edinburgh) at the Association for Machine Translation in the Americas Evaluation Workshop in San Diego, CA, in November, 1992, and further discussed in a subsequent personal communication from Margaret King (ISSCO, Geneva).

the information extraction tasks that have been used for MUC are quite realistic in some respects, and there are ways in which the evaluation metrics and scoring procedures reflect the concern that the interests of technology consumers be accomodated to the extent possible. Their interest is in adequacy evaluation, which is defined as "assessing the adequacy of a system with respect to some intended use of that system, as exemplified by a potential customer investigating whether a system, either in its current state or after modification, will do what he requires, how well it will do it and at what cost." The third type, diagnostic evaluation, is defined as "assessing the state of a system with the intention of discovering where it fails and why, as exemplified by a research group examining their own system." There are ways in which the MUC evaluations partially support this purpose as well, by providing quantitative data and by facilitating the collection of qualitative data.

4.1. Progress Evaluation

There are at least three ways we look at progress: as an assessment of the current state of the art, as a measure of progress relative to the previous evaluation, and as a measure of progress toward matching human performance on the same task. We expect the metrics to be applicable to both machines and humans, to provide a useful way to look at how much of the expected data the system is finding and at the classes and numbers of errors it is making, and to offer a means for comparing performance across systems.

Using the metrics that have been developed so far, we can say how systems are doing on particular information extraction tasks with respect to correct, incorrect, spurious and missing data at various levels of granularity, and we can tell how a system's performance on the parts of the task that it tried to do compares to its performance on the total task. Repeated over time, the assessments measure progress of the systems as a group and as individuals, although precise measurement has been complicated by the changes to the evaluation methodology, task domain, and template design, and by the radical system design changes made by some groups. Overall cross-system comparisons are possible given a single-value metric [2] and statistical significance tests [3]. The most compelling research problems posed by the task, e.g., suprasentential processing [7], are dramatically revealed.

In the context of ARPA's Tipster program, human performance studies have been carried out with the analysts who filled the answer-key templates. One of these studies [13], which was conducted in the English joint ventures domain, used 20 templates generated independently by four analysts and compared with a key prepared by a fifth "expert" partly on the basis of the other four. The results showed that the best performance achieved was 82% recall and 84% precision, that a fairly small amount of variability existed between the two top-scoring humans, and that there was a sizable performance difference between the top-scoring and the lowest-scoring humans.

An error analysis of these results showed that about half of the approximately 20% total disagreement among the analysts could be attributed to human error (misinterpretation, oversight, data-entry error). The rest was attributed to problems outside the human's control (gaps in template-filling guidelines, legitimate analytical differences in text and guideline interpretation, and bugs in the template-filling tool). Although human performance in this study is far from perfect, it nonetheless represents a challenging performance objective for computer systems.

4.2. Adequacy Evaluation

Although the evaluation tasks emulate actual or hypothesized real-life tasks, they are unrealistic in certain crucial respects, such as the complete autonomy of the extraction process. Since the tasks are constrained in ways such as this for the purposes of evaluation, it is not possible to translate the evaluation results directly into terms that reflect the specific requirements of any particular real-life applications, even applications that bear strong resemblances to the evaluation tasks. Nonetheless, we can consider the relevance of the MUC evaluation methodology to the problem of assessing the adequacy of systems and methods for real-life tasks.

Decisions concerning choice of evaluation metrics have been motivated in part by an interest in establishing good communications with technology consumers. As communications have improved, misconceptions concerning the presumed needs of technology consumers in terms of evaluation metrics have surfaced and are being addressed. The result should be a small set of easily-understood metrics that provide insightful performance data for consumers as well as producers.

One example concerns the treatment of missing and spurious fills, which has been left as a variable so that technology consumers can decide to what extent they are concerned with absent or excess data in the database. However, it now appears that a strict and equal treatment of both types of error is more meaningful to the technology consumers as well as to the technology producers. Another example concerns the overall metric that is computed primarily to enable systems to be ranked. The current metric was designed with the presumed interests of technology consumers in mind, by incorporating variable weights for recall and precision and

by including a factor that rewards systems for balanced performance on those two measures. However, there is strong interest among some technology users and others in replacing the current metric with the error rate (number wrong divided by total possible).

In addition to influencing the development of evaluation metrics, the concerns of adequacy evaluation have affected some of the decisions programmed into the scoring software. All in all, the MUC evaluations have quite consciously responded to some of the presumed needs of technology consumers; it now appears that one of our priorities should be to eliminate some of the embellishments and complexities that have been introduced over the last few years.

4.3. Diagnostic Evaluation

The primary metrics of recall and precision and the secondary ones of undergeneration and overgeneration provide diagnostic information in the sense that they show how accurate system performance is at the system's current level of task coverage. We rely on the evaluation participants for error analyses and qualitative assessments of their system's performance, using the metrics as one starting point. Attempts that have been made to use the information extraction task to reveal language analysis capabilities directly have so far met with limited success. Although these attempts have stayed within the "black-box" information extraction evaluation paradigm by examining only textual inputs in relation to template-filler outputs, they are diagnostic evaluations in the sense that they seek to isolate specific aspects of text analysis from the information extraction task, making use of test suites of examples selected from the overall extraction task.

One of the studies 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 [1] and the latter for MUC-4 [6]. In both studies, there were conditions where the results conformed to expectations and conditions where they did not. Both studies suffered from small test suites and a number of uncontrolled variables. 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* to the basic evaluation but rather require independent specification in order to ensure adequate test samples and an appropriately designed information extraction task.

5. CHALLENGES FOR THE FUTURE

A major challenge for the immediate future of the MUC evaluations is to make the results more intuitively meaningful and more directly usable by the various interested parties -- those doing the research and development, those watching, and those contemplating use. To date, the results seem to have served those doing the research and development well and the others not so well. Of benefit to all, however, have been the development of the shared tasks and the large prototype systems, which have provided the basis for effective communication.

The pressures of the information extraction evaluation tasks and the pressures of the evaluations themselves have resulted in increased attention to task-specific processing techniques. These techniques are often designed not only to improve the quantity and quality of extracted information but also to shorten the development cycle and reduce the human effort associated with porting and extending the system. At the extreme end of the spectrum is a class of systems that exploit various shallow processing techniques. The performance objective of such systems is to at least come close to the estimated potential performance of an in-depth understanding system and to reach that level with much less time and effort. Thus, the contrasts in system design philosophy and system architecture have grown, and the foundation has been laid for an evaluation that could reveal a lot about the near-term transition potential of some technologies and about the strategies for addressing the significant, longer-term research issues associated with the information extraction task.

Although information extraction has served as an excellent vehicle for elucidating the application potential of current technology, its utility as a vehicle for focusing attention on solving the hard, general problems of natural language processing is not as great. Many insights have been gained into the nature of natural language processing 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. This is especially true of the many MUC participants with severely limited resources, but it is also true to some extent for those with more extensive resources, who may feel the pressure of competition for high performance more keenly. It is clearly of little use to anyone to ask a large number of research-oriented groups to productize their systems and fine-tune them to a particular domain, just for the purposes of evaluation. The challenge to play a role in solving the hard natural language processing problems is a challenge for the evaluators and participants alike.

ACKNOWLEDGEMENTS

The authors are especially indebted to the other members of the MUC-5 program committee: Sean Boisen, Lynn Carlson, Jim Cowie, Ralph Grishman, Jerry Hobbs, Joe McCarthy, Mary Ellen Okurowski, Boyan Onyshkevych, and Carl Weir. The authors' work is funded by ARPA/SISTO under ARPA order 6359.

REFERENCES

1. 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.

2. Chinchor, N., MUC-4 Evaluation Metrics, in *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, June, 1992, Morgan Kaufmann Publishers, pp. 22-29.

3. Chinchor, N., Statistical Significance of MUC-4 Results, in *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, June, 1992, Morgan Kaufmann Publishers, pp. 30-50.

4. Chinchor, N., Hirschman, L., and Lewis, D.D., Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3), to appear in *Computational Linguistics*, 19(3).

5. 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 Publishers, pp. 25-30.

6. Hirschman, L., An Adjunct Test for Discourse Processing in MUC-4, in *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, June, 1992, Morgan Kaufmann Publishers, pp. 67-84.

7. Iwanska, L., et al., Computational Aspects of Discourse in the Context of MUC-3, in *Proceedings of the Third Message Understanding Conference (MUC-3)*, May, 1991, Morgan Kaufmann Publishers, pp. 256-282.

8. *Proceedings of the Third Message Understanding Conference (MUC-3)*, May, 1991, Morgan Kaufmann Publishers.

9. *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, June, 1992, Morgan Kaufmann Publishers.

10. 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 Publishers, pp. 197-202.

11. Sundheim, B., Third Message Understanding Evaluation and Conference (MUC-3): Phase 1 Status Report, in *Proceedings of the Speech and Natural Language Workshop*, February, 1991, Morgan Kaufmann Publishers, pp. 301-305.

12. Sundheim, B., Overview of the Fourth Message Understanding Evaluation and Conference, in *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, June, 1992, Morgan Kaufmann Publishers, pp. 3-21.

13. Will, C. and Onyshkevych, B., Human Performance for Information Extraction, unpublished presentation given at the Tipster 12-month meeting in San Diego, CA, September, 1992.