COMPARING MUCK-II AND MUC-3: ASSESSING THE DIFFICULTY OF DIFFERENT TASKS

Lynette Hirschman
Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, MA 02139
E-mail: hirschman@goldilocks.lcs.mit.edu

OVERVIEW

The natural language community has made impressive progress in evaluation over the last four years. However, as the evaluations become more sophisticated and more ambitious, a fundamental problem emerges: how to compare results across changing evaluation paradigms. When we change domain, task, and scoring procedures, as has been the case from MUCK-I to MUCK-II to MUC-3, we lose comparability of results. This makes it difficult to determine whether the field has made progress since the last evaluation. Part of the success of the MUC conferences has been due to the incremental approach taken to system evaluation. Over the four year period of the three conferences, the domain has become more "realistic", the task has become more ambitious and specified in much greater detail, and the scoring procedures have evolved to provide a largely automated scoring mechanism. This process has been critical to demonstrating the utility of the overall evaluation process. However we still need some way to assess overall progress of the field, and thus we need to compare results and task difficulty of MUC-3 relative to MUCK-II.

This comparison is complicated by the absence of any generally agreed upon metrics for comparing the difficulty of two natural language tasks. There is no real analog, for example, to *perplexity* in the speech community, which provides a rough cross-task assessment of the difficulty of recognizing speech, given a corpus and some grammar or language model for that corpus. Natural language does not have a set of metrics whose affect on task difficulty is well-understood. In the absence of such metrics, this paper outlines a set of dimensions that capture some of the differences between the two tasks. It remains a subject for further research to define appropriate metrics for cross-task comparison and to determine how these metrics correlate with performance.

Clearly it is impossible to come up with a single number that characterizes the relative difficulty of MUCK-II and MUC-3. Nonetheless, we can characterize both qualitative and quantitative differences along the following dimensions:

- Complexity of data
- Corpus dimensions
- Nature of the task
- Difficulty of the task
- Scoring of results

In general, it is safe to say that MUC-3 was much harder than MUCK-II in all of these dimensions. It is also clear that the scores for MUC-3 were lower (even after adjusting for the difference in how scores

Training Data	MUCK-II	MUC-3	FACTOR
No. Msg Types	4	16	4x
Vocabulary Size	1,899	18,240	10x
Ave. Sent Length*	12	27	2x
Ave. Msg. Length (in sentences)	3	12	4x

* Parse difficulty may increase as the cube of sentence length for some parsers

Table 1: Complexity of Data

were computed). The best precision and recall figures for MUCK-II were in the range of 70-80% on the final test (using a corrected score, calculated in the MUC-3 style). For MUC-3, precision and recall were in the range of 45-65% for the final test. Although the error rates were about double for MUC-3, the task was many times harder. From this, we can conclude that the field has indeed made impressive progress in the two years since MUCK-II.

DIMENSIONS OF COMPARISON

Complexity of the Data

The first respect in which MUCK-II and MUC-3 differ is in the type of messages chosen as input. MUCK-II used Navy message traffic (operational reports) as input, which are filled with jargon, but also cover a rather limited domain with a fairly small vocabulary (2000 words). These messages were characterized by a telegraphic syntax and somewhat random punctuation; run-on sentences, for example, occurred quite frequently. MUC-3 uses news wire reports as input, which are more general (in that they use less jargon), but cover a much wider range of subjects, with a much larger vocabulary (20,000 words). Although the syntax is generally more "standard" in MUC-3, the richness of the corpus poses new problems. Sentences can be extremely long and quite rich in their range of syntactic constructions (see Figure 1 for extreme examples of MUCK-II and MUC-3 sentences).

In MUCK-II, there were four distinct message types; in MUC-3, there are sixteen. These include "text", "communique" and "editorial" as well as translated transcriptions of radio communiques originally in Spanish. The average sentence length is longer in MUC-3 (27 words compared to 12 words for MUCK-II), as is message length (12 sentences/message for MUC-3 compared to 3 sentences/message for MUCK-II). These differences are summarized in Table 1.

None of the measures in Table 1 makes any attempt to measure grammatical complexity. Of the various measures, perhaps sentence length correlates most closely. In general, longer sentences are harder to parse; for certain types of parser, the time to parse increases as the cube of sentence length. Even this does not begin to describe the rich range of syntactic constructions in MUC-3. On the other hand, this richness is difficult to compare to the difficulties of handling telegraphic and heavily elliptical material in MUCK-II. Figure 1 contains a sentence from MUCK-II and a sentence from MUC-3, to illustrate the different problems posed by the two corpora.

In addition to the metrics reported in Table 1, there are other metrics which would help to capture the notion of data complexity. We could measure the rate of growth of the vocabulary, for example, to determine how frequently new words appear; this would also give us insight into whether we had a

MUCK-II

SEATTLE TAKEN UNDER FIRE BY KRESTAI, FRIENDLY FORCES AIR CONDITION WARNING REF WEAPON FREE ON HOSTILE SURFACE

MUC-3

He disclosed that, in August 1988, he was assigned the missions of planning the assassination of the president of the republic and the beheading of the Honduran spiritual and political leadership through the physical elimination of Honduran Archbishop Msgr Hector Enrique Santos and presidential candidates Carlo Roberto Flores and Rafael Leonardo Callejas Romero, of the Liberal and National parties respectively.

Figure 1: Sample Hard Sentences from MUCK-II and MUC-3

sufficient amount of training data. We could measure perplexity using a simple grammar (e.g., a bigram language model); this is an information theoretic measure which is standard in the speech recognition community and gives an idea of how many words (on average) can follow a given word in the corpus. There are also some system-dependent measures which would be interesting: number of grammar rules used, number of co-occurrence patterns, number of inference rules, and size of the knowledge base. Investigation of what to measure and how these measures relate to other things (e.g., accuracy, speed) should be considered a subject of further research in its own right, if we expect to make meaningful comparisons across different domains.

Corpus Dimensions

The size of the MUC-3 corpus has had a profound impact on how the participating message understanding systems were built and debugged. MUC-3 provided one to two orders of magnitude more data than MUCK-II: 1300 training messages and some 400,000 words, compared to 105 messages and 3,000 words for MUCK-II. If hand-debugging was still (barely) possible in MUCK-II, it was clearly an overwhelming task for MUC-3. In addition, system throughput became a major consideration in MUC-3. If it takes a system one day to process 100 messages, then running through the development corpus (1300 messages) would take two weeks – a serious impediment to system development. This has placed a greater premium on system throughput and on automated procedures for detecting errors and evaluating overall performance. It has also led some systems to explore methods for skimming and for distinguishing "important" or high-information passages from less important passages, as well as robust or partial parsing techniques.

The corpus dimensions differed for the test sets as well. The test set for MUC-3 consisted of 100 messages (33,000 words), compared to five messages for MUCK-II (158 words). The larger training corpus placed heavier processing demands on the systems, but it also meant that the test data was more representative of the training data. In MUCK-II, 16% of the total number of words in the test represented previously unseen tokens. In MUC-3, this figure dropped to 1.6%.

Nature of the Task

There was a slight change in task focus between MUCK-II and MUC-3. In MUCK-II, the task was template-fill, where each message generated at least one template (although one type of template was an "OTHER" template, indicating that it did not describe any significant event). All messages were

Training Data	MUCK-II	MUC-3	FACTOR
No.Texts	105	1,300	12x
No. Sentences	263	15,600	60x
No. Total Words	3,058	400,000	130x

Test Data	MUCK-II	MUC-3	FACTOR
No.Texts	5	100	20x
No. Sentences	18	1,183	66x
No. Total Words	158	33,615	200x
No. New Words*	26 (16%)	555 (1.6%)	0.1x

* Percent = new words over total number of words (tokens)

Table 2: Corpus Dimensions

considered to "relevant" in the sense of generating a template; only 5% of the training messages generated an "OTHER" template. MUC-3 required both relevance assessment and template fill: only about 50% of the messages were relevant. Part of the task involved determining whether a message contained relevant information, according to a complicated set of rules that distinguished current terrorist events from military attacks and from reports on no-longer-current terrorist events. Thus relevance assessment turned out to be a complex task, requiring four pages of instructions and definitions in MUC-3 (compared to half a page of instructions for MUCK-II). Filling a template for an irrelevant message was penalized – to a greater or lesser extent, depending on which metrics were used in the summary scoring.

Although this represents a change between the two tasks, it is difficult to come up with any numerical measures to quantify this difference. On the one hand, the participants reported that this shift did not contribute substantially to the difficulty. On the other hand, most sites devoted substantial effort to creating a specialized set of rules to distinguish relevant from irrelevant messages. Understanding these rules for relevance was certainly one of the least portable and most domain-specific aspects of the task, so it undoubtedly did contribute to the greater difficulty of MUC-3.

Difficulty of the Template Fill Task

Reflecting the change in application domains, the templates changed from MUCK-II to MUC-3. The templates differ in how many types of template there are, in the number of slots, in the allowable range of slot fills, and in number of fills per slot (since more than one fill is required in certain cases). MUCK-II had 6 types of events and 10 slots per template, of which five were filled from small closed class lists, three from larger closed class lists, and two by string fills from the text. MUC-3 had 10 types of events and 17 slots (not counting slots reserved for indexing messages and templates), of which eight were small closed classes, two were larger closed classes, and seven required numerical or string fills from the text.

For the MUCK-II test, there were 5 templates generated for 5 messages with just over one fill per slot (55 fills for 50 slots). For the MUC-3 test of 100 messages, 65 out of 100 were relevant. For the relevant messages, there were 133 templates generated (roughly 2 templates per message, counting the 19 optional templates). The ratio of slots to slot fillers was approximately 2500 answers for 2260 slots (1.1 answers/slot). However, many answers in MUC-3 included cross-references to other slot fillers, which were required to get full credit for a correct answer. There were approximately 1000 of these cross-references, so a more realistic estimate of number of "fills" was 3500 (1.5 answers/slot). This information

	MUCK-II	MUC-3	FACTOR
No. Template Types	6	10	2x
No. Slots	10	17	2x
Percent Relevant	100	65	-
No. Templates/Msg	1.1	2	2x
Answers/Slot*	1.1	1.5	1.4x
Types of Fill			
a. No. Fixed Fill < 10	5	8	
b. No. Fixed Fill < 100	3	2	
c. No. Numerical	0	3	
d. No. String Fill	2	4	
Overall Difficulty**	17	30	2x

* Answers for MUC-3 include cross-reference answers
** Difficulty = "perplexity" counting number of possible fills for a slot as the branching factor at that point

Table 3: Comparison of the Template Fill Task

is summarized in Table 3.

It is possible to enumerate the ways in which the two template fill tasks differ, but it is extremely difficult to assess how this affects the overall difficulty of filling the template. One crude approach is to compute a perplexity-like measure of the tasks, looking at the filled template as a string of answers, using the number of possible fills for each slot is an estimate of the branching factor at that point. This yields a "difficulty" figure of 17 for MUCK-II as opposed to a figure of 30 for MUC-3. This number corresponded to the perceived increase in difficulty between the two tasks by the participants: MUC-3 was definitely viewed as harder, but not an order of magnitude harder.

Scoring and Results

Finally, the two tasks also scored the results differently. MUCK-II generally used a score based around 1: wrong = 0, no answer = 1, right = 2. In MUC-3, the correct answer counted 1, the wrong answer counted 0. It is possible to recompute the scores for MUC-2 to make them comparable to MUC-3. If we do this, we find that the top-scoring systems in MUCK-II had precision and recall scores in the 70-80% range. This compares to 45-65% for the top-scoring systems in MUC-3 for the run which maximized both precision and recall using the "MATCHED-MISSING" method of computing the score[1]. Since 100% is the upper bound, it is actually more meaningful to compare the "shortfall" from the upper bound; for MUCK-II, this is 20-30% and for MUC-3, 35-55%. Thus MUC-3 performance is about half as good as (has twice the shortfall) as MUCK-II.

CONCLUSIONS

Table 4 attempts to summarize this discussion by providing a rough order of magnitude for the different dimensions. We see from this that MUC-3 is many times harder than MUCK-II, in three of

[1] The term "MATCHED-MISSING" refers to the metric which penalized systems for each missing template, but counted spurious templates wrong only in the template ID slot, not in each individual filled slot

DIMENSION	FACTOR
1. Complexity of Data	10x
2. Corpus Dimensions	100x
3. Nature of Task	-
4. Difficulty of Template Fill	2x
5. Overall Performance	0.5x

Table 4: Summary of Differences: MUCK-II vs. MUC-3

the four dimensions, while performance has only been cut by a factor of two. Even though the relation between difficulty and precision/recall figures is certainly not linear (the last 10-20% is always much harder to get than the first 80%), the degree of difficulty has increased much more than the performance has deteriorated.

This comparison is reassuring in several respects. First, it means that the field has made very substantial progress in the past two years. MUC-3 shows that current message understanding systems are able to handle a realistic corpus, with a realistic throughput with a reasonable degree of accuracy – higher precision and recall than many information retrieval systems are likely to get. Secondly, it means that as a test, MUC-3 was well-designed. Part of the motivation in changing tasks and domains after MUCK-II was to make the problem realistic and sufficiently challenging so that there would be no easy or trick solutions. MUC-3 has served that purpose admirably. It is realistic but current systems can achieve a reasonable level of performance. It is hard enough so that there is substantial room for improvement. This task can provide a reasonable challenge for message understanding systems over the next several years.

Finally, this comparison leads to an important conclusion about evaluation methodology. This paper represents a tentative first step towards defining some ways of measuring the dimensions of an application. But it is clear that we need to do much more work in this area in order to gain insight into what dimensions really affect success and which ones are less critical. We need to run experiments, where we can vary one set of parameters, while holding others constant. In short, to gain the maximum benefit from evaluation efforts such as the MUC conferences, we need to make evaluation methodology itself a legitimate topic for research.

ACKNOWLEDGEMENTS

I would like to thank a number of people for their help in assembling the data and providing insights into this problem. In particular, I would like to acknowledge the help that I received from Beth Sundheim at NOSC, who furnished me with much of the data and some very helpful suggestions for possible ways to compare the MUCK-II and MUC-3 tasks. I would also like to thank Ralph Grishman (NYU), Jerry Hobbs (SRI), Wendy Lehnert (University of Massachusetts) and Lisa Rau (GE) for feedback and suggestions. Any errors or controversial conclusions are, however, my sole responsibility.