MITRE-Bedford ALEMBIC: MUC-4 Test Results and Analysis

John Aberdeen, John Burger, Dennis Connolly, Susan Roberts, & Marc Vilain

{*aberdeen, john, decon, suzi, mbv*}*@mitre.org*

The MITRE Corporation
202 Burlington Road
Bedford, MA 01730

PRELIMINARIES

This note embodies our analyses of the performance of the ALEMBIC system in the MUC-4 evaluation task. These analyses have provided us with a reasonably good understanding of the principal factors contributing to the system's correct responses and to its errors. This understanding is based in part on interpretations of the performance measures provided by the MUC-4 scoring software; in addition, we performed a number of qualitative and quantitative investigations into linguistic aspects of the messages that underly the system's performance.

It should be noted however, that ALEMBIC is still in very early stages of development, and that the analyses we give here should be taken as just presenting a snapshot of the system's performance. In the weeks since the MUC-4 evaluation runs, the system has of course remained under development, and its performance scores have improved steadily. One consequence of our system's relative youth is that it embodies many oportunities for improvement, and even minor implementational tweaks can yield significant performance gains.

OVERALL PERFORMANCE MEASURES

Looking first at our system's overall performance, the following table reproduces the f-measures for our runs on TST3 and TST4.

	P&R	2P&R	P&2R
TST3	9.6	8.57	10.91
TST4	13.75	11.22	17.74

Table 1: Overall f-scores

These are clearly fairly humble scores, but we offer them for consideration with a certain measure of pride. They represent the very first results of a text understanding project that was barely begun six months prior to the evaluation runs, as fielded by a group that had no prior experience with the MUC data extraction task.

As one might note, raw f-measures are a fairly coarse performance statistic; despite this, some trends are apparent. First, our system seems better at recall than at precision, an issue which we will address below. Second, the system scored uniformly better on TST4 than on TST3, which is in line with the general observation among MUC-4 participants that TST4 is the easier of the two test sets.

RECALL MEASURES

Turning to a more detailed analysis of our recall measures, the principal determinant of our overall recall performance is rather clear. That is, we only attempted to fill about half of all possible template slots, those for the incident and perpetrator. Although we had slot-filling strategies prepared for the remaining slots, they were never incorporated into the system prior to the evaluation runs—we just simply ran out of time to do so.

It is illuminating, however, to consider ALEMBIC's performance on the slots that the system actually did fill. The more interesting of these slots are those that the system filled by meaningfully extracting information from the free text; their slotwise recall scores are shown in Table 2 below. Slots that do not appear in the table were simply not filled at all, or were only filled by default strategies (more on this later).

	TST3	TST4
inc-date	23	35
inc-loc	10	16
inc-type	30	53
inc-instr-id	8	27
inc-instr-type	3	16
perp-ind-id	22	48
perp-org-id	40	40

Table 2: Recall scores on meaningfully filled slots

A quick glance at the table reveals that our scores ranged fairly widely. On string fills, ALEMBIC obtained scores ranging for TST3 from 8 (instrument ID) to 40 (perpetrator organization ID); for set fills the range was 3 (instrument type) to 30 (incident type). Similar patterns held for TST4, but with higher individual slot scores, reflecting the fact that this was the easier of the two test sets. As an estimate of the average recall for the slots in Table 2 , we calculated a restricted overall recall score (based only on these slots) of approximately 20 for TST3 and 34 for TST4.[1]

On a slot-by-slot basis, the following qualitative observations apply.

Incident date: We derived this slot from the free text, and only used the dateline as a last recourse in case we failed to identify any date phrases. The date grammar we used for MUC-4 treats date phrases as functors, which were often left unattached due to the fragmentary nature of our parses. This made it harder to actually locate temporal phrases when they did not appear as modifiers of events, resulting in a fair number of invocations of the heuristic fallback strategy of using the dateline.

Incident location: Recall errors for this slot were due in good part to locational modifiers not being attached to events, as well as to a number of infelicities in the locational knowledge representation. Among the more amusing: the lexical item *Bogota* maps to a number of possible locations, but the one that was picked by default was the Bogota Air Force Base.

Incident type: This was our most accurate set-fill slot. ALEMBIC derives the filler of this slot from the heads of violent events; missing cases are due in part to gaps in the lexicon.

[1]Not too much should be made of these scores. They admittedly exclude slots that are easy to fill using default values, but they also don't include slots that are hard to fill, i.e., the target slots.

Incident instrument ID: We expected to get better recall scores for this slot. Eyeballing the actual fillers that ALEMBIC produced, part of the problem was grammatical incompleteness. For example, "a powerful dynamite charge" ended up only being parsed as "a powerful dynamite," due to a grammar bug involving noun-noun modification. Since we attempted to use full noun phrases to fill string slots, we ended up being penalized for cases where we had nearly parsed the complete instrument phrase, but where our fragmentary filler failed to be matched by the scoring program.

Incident instrument type: This slot was only filled when an instrument ID filler was obtained. We never implemented implicit fills for this slot, i.e., fills that could be derived from verbs such as *shoot* even if no gun is ever mentioned.

Perpetrator individual ID: As mentioned in the system overview, our strategy for filling this slot was heuristic. In case the violent event associated with the template lacked an agentive argument, plausible candidates were looked for elsewhere in the neighboring text. Once again, the fragmentary nature of the parses led to the heuristic fallback strategy being invoked fairly often, with very mixed results.

Perpetrator organization ID: We obtained comparatievely high recall scores for this slot. This is a relatively easy slot to fill, however, because likely perpetrator organizations are readily identified.

The remaining incident and perpetrator slots ended up being filled by default values. As a result, although we obtained some reasonable recall scores for these individual slots, these scores are of little real interest.

ISSUES WITH PRECISION

Our precision error rate is largely accounted for by overly eager template generation. As we note in the system description, the version of ALEMBIC fielded at MUC-4 generates a template for every seemingly distinct violent event. Our strategy for distinguishing such events from each other was heavily dependent on our reference resolution module, which turned out to be quite unreliable, and as a result generated multiple (nearly) identical templates for the same event. Consequently, we ended up with overall low precision and high overgeneration scores, as demonstrated by Table 3.

	Precision	Overgeneration
TST3	8	90
TST4	10	87

Table 3: Overall precision scores (all templates row)

The effects of this template generation strategy on our precision scores were fairly dramatic. We ended up actually making relatively few incorrect fills for those templates that were mapped by the scoring program. Specifically, ignoring spurious templates (as in the matched/missing row) we obtain precision scores of 72 and 75 for TST3 and TST4 respectively. However, because our spurious fills ended up outnumbering our incorrect fills by 20 to 1, our official scores from the all templates row were considerably weaker.

Aside from this principal cause of our precision errors, another significant factor is that ALEMBIC failed to filter out templates that corresponded to military clashes between guerrilla groups and the armed forces. We failed to incorporate such a filter largely because to do so presupposes filling some slots that we were simply leaving blank. Anecdotally, among the worst offenders of this sort was one message for which we generated two legitimate templates and ten templates corresponding to military clashes.

A CLOSER LOOK AT SYNTAX AND REFERENCE

A common thread to both our recall and precision problems is fragmentation of the parses. With respect to recall, fragmentation lead to the syntactic arguments of event verbs (and of their nominalizations) being left unattached; this caused the system to fall back frequently on unreliable backup strategies. In addition, the fragmentation confused our reference resolution module, because it introduced far too many top-level noun phrases or event verbs, each of which was potentially a candidate for reference resolution. This caused ALEMBIC to miss co-references needed to fill slots, and it also led to the system's poor ability to distinguish identical events on the basis of reference resolution.

We performed a number of post-hoc analyses to estimate the relative weight on fragmentation of various linguistic factors for which the MUC-4 version of ALEMBIC had incomplete grammatical coverage. These turned out to include a traditional and unsurprising cast of linguistic characters: coordination, PP attachment, noun-noun modification, subgrammars for title, date, location, *etc*. None of these factors seems particularly dominant, however; they all need to be eventually addressed in some way (linguistically principled or otherwise).

Because fragmentation played such a compromising role with respect to our reference resolution module, we also performed a number of quantitative analyses to clarify the nature of the problem. To begin with, we looked at the set of candidates that were considered when resolving an anaphoric expression. Many of these candidates were spurious; they had been introduced by (for instance) failing to attach a premodifier noun to its head noun, thus generating two top-level discourse entities where there should have been one. To our surprise however, many of these candidates turned out to be not just spurious but actually indistinguishable. Indeed, the Bayesian reference mechanism operates on the basis of a number of attributes of the referring expression and the potential candidate, e.g., KR agreement, number agreement, and others. It turned out however, that by this set of distinguishing attributes, there were on average 4.8 indistinguishable candidates for each anaphoric referring expression (based on a test set of 100 such anaphoric expressions). For pronominal expressions in particular, there were 3.6 such candidates and for definite NP's there were 5.1.

Three comments bear on these results. First, reducing fragmentation is imperative before the reference resolution module in ALEMBIC can perform its intended task. Second, the Bayesian reference network needs to take into account a greater number of distinguishing criteria than it currently does. For example, the reference resolution literature has frequently noted that grammatical case (subject, object, *etc*) is a strog cue towards reference resolution—we currently fail to take this cue into account, in good part because our parses are too fragmentary to produce reliable grammatical case assignments. Finally, we should note that the Bayesian network was trained on relatively poor-quality data, namely the fragmentary output of the parser. We expect that as the parser improves overall, these training data will become more pointed, and the resolution mechanism will improve as a consequence.

CONCLUDIGN THOUGHTS

As with all other participants in MUC-4, the evaluation process has taught us much about our system, and given us the oportunity to examine critically many of our design decisions. We are still extremely happy with much of the ALEMBIC architecture, and with much of the actual implementation. We also regret not having focussed more on some aspects of the system. To name a few: head feature agreement (whose implementation would have saved us much grammar tweaking), event merging, and of course more comprehensive slot filling.MUC-

The MUC- process has also been valuable in another way. Having now been exposed to the details of other participating systems, we have also become aware of a host of clever tricks that we are now eager to try.

We look forward to MUC-5 as an an oportunity to prepare the system we really had wanted to field for MUC-4.