Message Understanding Conference - 6: A Brief History

Ralph Grishman
Dept. of Computer Science
New York University
715 Broadway, 7th Floor
New York, NY 10003, USA
grishman@cs.nyu.edu

Beth Sundheim
Naval Command, Control and
Ocean Surveillance Center
Research, Development, Test and
Evaluation Division (NRaD)
Code 44208
53140 Gatchell Road
San Diego, California 92152-7420
sundheim@pojke.nosc.mil

Abstract

We have recently completed the sixth in a series of "Message Understanding Conferences" which are designed to promote and evaluate research in information extraction. MUC-6 introduced several innovations over prior MUCs, most notably in the range of different tasks for which evaluations were conducted. We describe some of the motivations for the new format and briefly discuss some of the results of the evaluations.

1 The MUC Evaluations

We have just completed the sixth in a series of Message Understanding Conferences, which have been organized by NRAD, the RDT&E division of the Naval Command, Control and Ocean Surveillance Center (formerly NOSC, the Naval Ocean Systems Center) with the support of DARPA, the Defense Advanced Research Projects Agency. This paper looks briefly at the history of these Conferences and then examines the considerations which led to the structure of MUC-6.[1]

The Message Understanding Conferences were initiated by NOSC to assess and to foster research on the automated analysis of military messages containing textual information. Although called "conferences", the distinguishing characteristic of the MUCs are not the conferences themselves, but the evaluations to which participants must submit in order to be permitted to attend the conference. For each MUC, participating groups have been given sample messages and instructions on the type of information to be extracted, and have developed a system to process such messages. Then, shortly before the conference, participants are given a set of test messages to be run through their system (without making any changes to the system); the output of each participant's system is then evaluated against a manually-prepared answer key.

The MUCs are remarkable in part because of the degree to which these evaluations have defined a program of research and development. DARPA has a number of information science and technology programs which are driven in large part by regular evaluations. The MUCs are notable, however, in that they in large part have shaped the research program in information extraction and brought it to its current state.[2]

2 Early History

MUC-1 (1987) was basically exploratory; each group designed its own format for recording the information in the document, and there was no formal evaluation. By MUC-2 (1989), the task had crystalized as one of template filling. One receives a description of a class of events to be identified in the text; for each of these events one must fill a template with information about the event. The template has slots for information about the event, such as the type of event, the agent, the time and place, the effect, etc. For MUC-2, the template had 10 slots. Both MUC-1 and MUC-2 involved sanitized forms of military messages about naval sightings and engagements.

The second MUC also worked out the details of the primary evaluation measures, recall and precision. To present it in simplest terms, suppose the answer key has N_{key} filled slots; and that a

[1] The full proceedings of the conference are to be distributed by Morgan Kaufmann Publishers, San Mateo, California; earlier MUC proceedings, for MUC-3, 4, and 5, are also available from Morgan Kaufmann.

[2] There were, however, a number of individual research efforts in information extraction underway before the first MUC, including the work on information formatting of medical narrative by Sager at New York University; the formatting of naval equipment failure reports by Marsh at the Naval Research Laboratory; and the DBG work by Logicon for RADC.

system fills $N_{correct}$ slots correctly and $N_{incorrect}$ incorrectly (with some other slots possibly left unfilled). Then

$$\begin{aligned} recall &= \frac{N_{correct}}{N_{key}} \\ precision &= \frac{N_{correct}}{N_{correct} + N_{incorrect}} \end{aligned}$$

For MUC-3 (1991), the task shifted to reports of terrorist events in Central and South America, as reported in articles provided by the Foreign Broadcast Information Service, and the template became somewhat more complex (18 slots). This same task was used for MUC-4 (1992), with a further small increase in template complexity (24 slots).

MUC-5 (1993), which was conducted as part of the Tipster program,[3] represented a substantial further jump in task complexity. Two tasks were involved, international joint ventures and electronic circuit fabrication, in two languages, English and Japanese. The joint venture task required 11 templates with a total of 47 slots for the output — double the number of slots defined for MUC-4 — and the task documentation was over 40 pages long.

One innovation of MUC-5 was the use of a nested template structure. In earlier MUCs, each event had been represented as a single template – in effect, a single record in a data base, with a large number of attributes. This format proved awkward when an event had several participants (e.g., several victims of a terrorist attack) and one wanted to record a set of facts about each participant. This sort of information could be much more easily recorded in the hierarchical structure introduced for MUC-5, in which there was a single template for an event, which pointed to a list of templates, one for each participant in the event.[4]

3 MUC-6: initial goals

DARPA convened a meeting of Tipster participants and government representatives in December 1993 to define goals and tasks for MUC-6.[5] Among the goals which were identified were

- demonstrating task-independent component technologies of information extraction which would be immediately useful
- encouraging work to make information extraction systems more portable
- encouraging work on "deeper understanding"

Each of these can been seen in part as a reaction to the trends in the prior MUCs. The MUC-5 tasks, in particular, had been quite complex and a great effort had been invested by the government in preparing the training and test data and by the participants in adapting their systems for these tasks. Most participants worked on the tasks for 6 months; a few (the Tipster contractors) had been at work on the tasks for considerably longer. While the performance of some systems was quite impressive (the best got 57% recall, 64% precision overall, with 73% recall and 74% precision on the 4 "core" template types), the question naturally arose as to whether there were many applications for which an investment of one or several developers over half-a-year (or more) could be justified.

Furthermore, while so much effort had been expended, a large portion was specific to the particular tasks. It wasn't clear whether much progress was being made on the underlying technologies which would be needed for better understanding.

To address these goals, the meeting formulated an ambitious menu of tasks for MUC-6, with the idea that individual participants could choose a subset of these tasks. We consider the three goals in the three sections below, and describe the tasks which were developed to address each goal.

4 Short-term subtasks

The first goal was to identify, from the component technologies being developed for information extraction, functions which would be of practical use, would be largely domain independent, and could in the near term be performed automatically with high accuracy. To meet this goal the committee developed the "named entity" task, which basically involves identifying the names of all the people, organizations, and geographic locations in a text.

The final task specification, which also involved time, currency, and percentage expressions, used SGML markup to identify the names in a text. Figure 1 shows a sample sentence with named entity annotations. The tag **ENAMEX** ("entity name

[3] Tipster is a U.S. Government program of research and development in the areas of information retrieval and information extraction.

[4] In fact the MUC-5 structure was much more complex, because there were separate templates for products, time, activities of organizations, etc.

[5] The representatives of the research community were Jim Cowie, Ralph Grishman (committee chair), Jerry Hobbs, Paul Jacobs, Len Schubert, Carl Weir, and Ralph Weischedel. The government people attending were George Doddington, Donna Harman, Boyan Onyshkevych, John Prange, Bill Schultheis, and Beth Sundheim.

```
Mr.  <ENAMEX TYPE="PERSON">Dooner</ENAMEX> met with <ENAMEX TYPE="PERSON">Martin
Puris</ENAMEX>, president and chief executive officer of <ENAMEX
TYPE="ORGANIZATION">Ammirati & Puris</ENAMEX>, about <ENAMEX
TYPE="ORGANIZATION">McCann</ENAMEX>´s acquiring the agency with billings of <NUMEX
TYPE="MONEY">$400 million</NUMEX>, but nothing has materialized.
```

Figure 1: Sample named entity annotation.

expression") is used for both people and organization names; the tag NUMEX ("numeric expression") is used for currency and percentages.

5 Portability

The second goal was to focus on portability in the information extraction task — the ability to rapidly retarget a system to extract information about a different class of events. The committee felt that it was important to demonstrate that useful extraction systems could be created in a few weeks. To meet this goal, we decided that the information extraction task for MUC-6 would have to involve a relatively simple template, more like MUC-2 than MUC-5; this was dubbed "mini-MUC". In keeping with the hierarchical template structure introduced in MUC-5, it was envisioned that the mini-MUC would have an event-level template pointing to templates representing the participants in the event (people, organizations, products, etc.), mediated perhaps by a "relational" level template.

To further increase portability, a proposal was made to standardize the lowest-level templates (for people, organizations, etc.), since these basic classes are involved in a wide variety of actions. In this way, MUC participants could develop code for these low-level templates once, and then use them with many different types of events. These low-level templates were named "template elements".

As the specification finally developed, the template element for organizations had six slots, for the maximal organization name, any aliases, the type, a descriptive noun phrase, the locale (most specific location), and country. Slots are filled only if information is explicitly given in the text (or, in the case of the country, can be inferred from an explicit locale). The text

> We are striving to have a strong renewed creative partnership with Coca-Cola," Mr. Dooner says. However, odds of that happening are slim since word from Coke headquarters in Atlanta is that...

would yield an organization template element with five of these six slots filled:

```
<ORGANIZATION-9402240133-5> :=
    ORG_NAME: "Coca-Cola"
    ORG_ALIAS: "Coke"
    ORG_TYPE: COMPANY
    ORG_LOCALE: Atlanta CITY
    ORG_COUNTRY: United States
```

(the first line identifies this as organization template 5 from article 9402240133).

Ever on the lookout for additional evaluation measures, the committee decided to make the creation of template elements for all the people and organizations in a text a separate MUC task. Like the named entity task, this was also seen as a potential demonstration of the ability of systems to perform a useful, relatively domain independent task with near-term extraction technology (although it was recognized as being more difficult than named entity, since it required merging information from several places in the text). The old-style MUC information extraction task, based on a description of a particular class of events (a "scenario") was called the "scenario template" task. A sample scenario template is shown in the appendix.

6 Measures of deep understanding

Another concern which was noted about the MUCs is that the systems were tending towards relatively shallow understanding techniques (based primarily on local pattern matching), and that not enough work was being done to build up the mechanisms needed for deeper understanding. Therefore, the committee, with strong encouragement from DARPA, included three MUC tasks which were intended to measure aspects of the internal processing of an information extraction or language understanding system. These three tasks, which were collectively called SemEval ("Semantic Evaluation") were:

- **Coreference**: the system would have to mark coreferential noun phrases (the initial specification envisioned marking set-subset and part-whole relations, in addition to identity relations)
- **Word sense disambiguation**: for each open class word (noun, verb, adjective, adverb) in the text, the system would have to

determine its sense using the Wordnet classification (its "synset", in Wordnet terminology)

- **Predicate-argument structure**: the system would have to create a tree interrelating the constituents of the sentence, using some set of grammatical functional relations

The committee recognized that, in selecting such internal measures, it was making some presumption regarding the structures and decisions which an analyzer should make in understanding a document. Not everyone would share these presumptions, but participants in the next MUC would be free to enter the information extraction evaluation and skip some or all of these internal evaluations. Language understanding technology might develop in ways very different from those imagined by the committee, and these internal evaluations might turn out to be irrelevant distractions. However, from the current perspective of most of the committee, these seemed fairly basic aspects of understanding, and so an experiment in evaluating them (and encouraging improvement in them) would be worthwhile.

7 Preparation process

Round 1: Resolution of SemEval

The committee had proposed a very ambitious program of evaluations. We now had to reduce these proposals to detailed specifications. The first step was to do some manual text annotation for the four tasks — named entity and the SemEval triad — which were quite different from what had been tried before. Brief specifications were prepared for each task, and in the spring of 1994 a group of volunteers (mostly veterans of earlier MUCs) annotated a short newspaper article using each set of specifications.

Problems arose with each of the SemEval tasks.

- For coreference, there were problems identifying part-whole and set-subset relations, and distinguishing the two; a decision was later made to limit ourselves to identity relations.

- For sense tagging, the annotators found that in some cases Wordnet made very fine distinctions and that making these distinctions consistently in tagging was very difficult.

- For predicate-argument structure, practically every new construct beyond simple clauses and noun phrases raised new issues which had to be collectively resolved.

Beyond these individual problems, it was felt that the menu was simply too ambitious, and that we would do better by concentrating on one element of the Semeval triad for MUC-6; at a meeting held in June 1994, a decision was made to go with coreference. In part, this reflected a feeling that the problems with the coreference specification were the most amenable to solution. It also reflected a conviction that coreference identification had been, and would remain, critical to success in information extraction, and so it was important to encourage advances in coreference. In contrast, most extraction systems did not build full predicate-argument structures, and word-sense disambiguation played a relatively small role in extraction (particularly since extraction systems operated in a narrow domain).

The coreference task, like the named entity task, was annotated using SGML notation. A `COREF` tag has an `ID` attribute which identifies the tagged noun phrase or pronoun. It may also have an attribute of the form `REF`=n, which indicates that this phrase is coreferential with the phrase with ID n. Figure 2 shows an excerpt from an article, annotated for coreference.[6]

Round 2: annotation

The next step was the preparation of a substantial training corpus for the two novel tasks which remained (named entity and coreference). SRA Corporation kindly provided tools which aided in the annotation process. Again a stalwart group of volunteer annotators was assembled;[7] each was provided with 25 articles from the Wall Street Journal. There was some overlap between the articles assigned, so that we could measure the consistency of annotation between sites. This annotation was done in the winter of 1994-95.

A major role of the annotation process was to identify and resolve problems with the task specifications. For named entities, this was relatively straightforward. For coreference, it proved remarkably difficult to formulate guidelines which were reasonably complete and consistent.[8]

[6] The `TYPE` and `MIN` attributes which appear in the actual annotation have been omitted here for the sake of readability.

[7] The annotation groups were from BBN, Brandeis Univ., the Univ. of Durham, Lockheed-Martin, New Mexico State Univ., NRaD, New York Univ., PRC, the Univ. of Pennsylvania, SAIC (San Diego), SRA, SRI, the Univ. of Sheffield, Southern Methodist Univ., and Unisys.

[8] As experienced computational linguists, we probably should have known better than to think this was an easy task.

```
Maybe <COREF ID="136" REF="134">he</COREF>´ll even leave something from <COREF
ID="138" REF="139"><COREF ID="137" REF="136">his</COREF> office</COREF> for <COREF
ID="140" REF="91">Mr. Dooner</COREF>. Perhaps <COREF ID="144">a framed page from
the New York Times, dated Dec. 8, 1987, showing a year-end chart of the stock market
crash earlier that year</COREF>. <COREF ID="141" REF="137">Mr. James</COREF> says
<COREF ID="142" REF="141">he</COREF> framed <COREF ID="143" REF="144"
STATUS="OPT">it</COREF> and kept <COREF ID="145" REF="144">it</COREF> by <COREF
ID="146" REF="142">his</COREF> desk as a "personal reminder. It can all be gone like
that."
```

Figure 2: Sample coreference annotation.

Round 3: dry run

Once the task specifications seemed reasonably stable, NRaD organized a "dry run" – a full-scale rehearsal for MUC-6, but with all results reported anonymously. The dry run took place in April 1995, with a scenario involving labor union contract negotiations. Of the sites which were involved in the annotation process, ten participated in the dry run. Results of the dry run were reported at the Tipster Phase II 12-month meeting in May 1995.

8 The formal evaluation

The MUC-6 formal evaluation was held in September 1995. The scenario definition was distributed at the beginning of September; the test data was distributed four weeks later, with results due by the end of the week. The scenario involved changes in corporate executive management personnel. The evaluation met many of the goals which had been set by the initial planning conference in December of 1993.

There were evaluations for four tasks: named entity, coreference, template element, and scenario template. There were 16 participants; 15 participated in the named entity task, 7 in coreference, 11 in template element, and 9 in scenario template.

Named entity was intended to be a simple task on which systems could demonstrate a high level of performance ... high enough for immediate use. Our success in this task exceeded our expectations. The majority of sites had recall and precision over 90%; the highest-scoring system had a recall of 96% and a precision of 97%. Although one must keep in mind the somewhat limited range of texts in the test set (all are from the Wall Street Journal, in particular), the results are excellent. A couple of these systems have been commercialized, and several are being incorporated into government text-processing systems. Given this level of performance, there is probably little point in repeating this task with the same ground rules in a future MUC (although there might be interest in processing monocase text and in performing comparable tasks on a more varied corpus and for languages other than English).

The **template element** task, while superficially similar to named entities — it is also based on identifying people and organizations — is significantly more difficult. One has to identify descriptions of entities ("a distributor of kumquats") as well as names. If an entity is mentioned several times, possibly using descriptions or different forms of a name, these need to be identified together; there should be only one template element for each entity in an article. Consequently, the scores were appreciably lower, ranging across most systems from 65 to 75% in recall, and from 75% to 85% in precision. The top-scoring system had 75% recall, 86% precision. Systems did particularly poorly in identifying descriptions; the highest-scoring system had 38% recall and 51% precision for descriptions.

There seemed general agreement that having prepared code for template elements in advance did make it easier to port a system to a new scenario in a few weeks. This factor, and the room that exists for improvement in performance, suggest that including this task in a future MUC may be worthwhile.

The goal for **scenario templates** — mini-MUC — was to demonstrate that effective information extraction systems could be created in a few weeks. This too was successful. Although it is difficult to meaningfully compare results on different scenarios, the scores obtained by most systems after a few weeks (40% to 50% recall, 60% to 70% precision) were comparable to the best scores obtained in prior MUCs. The highest performance overall was 47% recall and 70% precision.

One can observe an increasing convergence of methods for information extraction. Most of the systems participating in MUC-6 employed a cascade of finite-state pattern recognizers, with the earlier pattern sets recognizing entities, and the later sets recognizing scenario-specific patterns.

This convergence may be one reason for the bunching of scores for this task — most systems fell in a rather narrow range in both recall and precision.

The results of this MUC provide valuable positive testimony on behalf of information extraction, but further improvement in both portability and performance is needed for many applications. With respect to portability, customers would like to have systems which can be ported in a few hours, or at most a few days, by someone with less expertise than a system developer. How this might be tested in the context of a MUC is not entirely clear. For one thing, most sites spent several days just studying the scenario description and annotated corpus, in order to understand the scenario definition, before coding began. Perhaps a micro-MUC[9] with an even simpler template structure, is needed to push the limits of portability. Getting systems which can be customized by others is also a tall order, given the complexity and variety of knowledge sources needed for a typical MUC information extraction task.

With respect to performance, the bunching of scores suggests that many sites were able to solve a common set of "easy" problems, but were stymied in processing messages which involved "hard" problems. Whether this is true, and just what the hard problems are, will require more extensive analysis of the results of MUC-6. Are the shortcomings due primarily to a lack of coverage in the basic patterns, to a lack of background knowledge in the domain, to failures in coreference, or something else? We may hope that the failings are primarily in one area, so that we may concentrate our energies there, but more likely the failings will be in many areas, and broad improvements in extraction engines will be needed to improve performance.

Pushing improvements in the underlying technology was one of the goals of SemEval and its current survivor, **coreference**. Much of the energy for the current round, however, went into honing the definition of the task. Philosophers of language have been arguing over reference and coreference for centuries, so we should not have been surprised that it would be so hard to prepare a precise and consistent definition. Additional work on the definition will be necessary, and it may be necessary to narrow the task further. Despite these distractions, a few interesting early results were obtained regarding coreference methods; we may hope that, once the task specification settles down, the availability of coreference-annotated corpora and the chance for glory in further evaluations will encourage more work in this area.

[9] a term suggested by George Krupka

Appendix: Sample Scenario Template

Shown below is a set of templates for the MUC-6 scenario template task. The scenario involved changes in corporate executive management personnel. For the text

> McCann has initiated a new so-called global collaborative system, composed of world-wide account directors paired with creative partners. In addition, Peter Kim was hired from WPP Group's J. Walter Thompson last September as vice chairman, chief strategy officer, world-wide.

the following templates were to be generated:

```
<SUCCESSION_EVENT-9402240133-3> :=
    SUCCESSION_ORG: <ORGANIZATION-9402240133-1>
    POST: "vice chairman, chief strategy
           officer,  world-wide"
    IN_AND_OUT: <IN_AND_OUT-9402240133-5>
    VACANCY_REASON: OTH_UNK
<IN_AND_OUT-9402240133-5> :=
    IO_PERSON: <PERSON-9402240133-5>
    NEW_STATUS: IN
    ON_THE_JOB: YES
    OTHER_ORG: <ORGANIZATION-9402240133-8>
    REL_OTHER_ORG: OUTSIDE_ORG
<ORGANIZATION-9402240133-1> :=
    ORG_NAME: "McCann"
    ORG_TYPE: COMPANY
<ORGANIZATION-9402240133-8> :=
    ORG_NAME: "J. Walter Thompson"
    ORG_TYPE: COMPANY
<PERSON-9402240133-5> :=
    PER_NAME: "Peter Kim"
```

Although we cannot explain all the details of the template here, a few highlights should be noted. For each executive post, one generates a SUCCESSION_EVENT template, which contains references to the ORGANIZATION template for the organization involved, and the IN_AND_OUT template for the activity involving that post (if an article describes a person leaving and a person starting the same job, there will be two IN_AND_OUT templates). The IN_AND_OUT template contains references to the templates for the PERSON and for the ORGANIZATION from which the person came (if he/she is starting a new job). The PERSON and ORGANIZATION templates are the "template element" templates, which are invariant across scenarios.