OVERVIEW OF RESULTS OF THE MUC-6 EVALUATION

Beth M. Sundheim

Naval Command, Control, and Ocean Surveillance Center
RDT&E Division (NRaD)
Information Access Technology Project Team, Code 44208
San Diego, CA 92152-7420
sundheim@nosc.mil

INTRODUCTION

The latest in a series of natural language processing system evaluations was concluded in October 1995 and was the topic of the Sixth Message Understanding Conference (MUC-6) in November. Participants were invited to enter their systems in as many as four different task-oriented evaluations. The Named Entity and Coreference tasks entailed Standard Generalized Markup Language (SGML) annotation of texts and were being conducted for the first time. The other two tasks, Template Element and Scenario Template, were information extraction tasks that followed on from the MUC evaluations conducted in previous years. The evolution and design of the MUC-6 evaluation are discussed in the paper by Grishman and Sundheim in this volume. All except the Scenario Template task are defined independently of any particular domain.

This paper surveys the results of the evaluation on each task and, to a more limited extent, across tasks. Discussion of the results for each task is organized generally under the following topics:

- Results on task as whole;
- Results on some aspects of task;
- Performance on "walkthrough article."

The walkthrough article is an article selected from the test set. Participants were asked to analyze their system's performance on that article and comment on it in their presentations and papers.

EVALUATION TASKS

A basic characterization of the challenge presented by each evaluation task is as follows:

- **Named Entity (NE)** -- Insert SGML tags into the text to mark each string that represents a person, organization, or location name, or a date or time stamp, or a currency or percentage figure.
- **Coreference (CO)** -- Insert SGML tags into the text to link strings that represent coreferring noun phrases.
- **Template Element (TE)** -- Extract basic information related to organization and person entities, drawing evidence from anywhere in the text.
- **Scenario Template (ST)** -- Drawing evidence from anywhere in the text, extract prespecified event information, and relate the event information to the particular organization and person entities involved in the event.

The two SGML-based tasks required innovations to tie system-internal data structures to the original text so that the annotations could be inserted by the system without altering the original text in any other way. This capability has other useful applications as well, e.g., it enables text highlighting in a browser. It also facilitates information extraction, since some of the information in the extraction templates is in the form of literal text strings, which some systems have in the past had difficulty reproducing in their output.

The inclusion of four different tasks in the evaluation implicitly encouraged sites to design general-purpose architectures that allow the production of a variety of types of output from a single internal representation in order to allow use of the full range of analysis techniques for all tasks. Even the simplest of the tasks, Named Entity, occasionally requires in-depth processing, e.g., to determine whether "60 pounds" is an expression of weight or of monetary value. Nearly half the sites chose to participate in all four tasks, and all but one site participated in at least one SGML task and one extraction task.

The variety of tasks designed for MUC-6 reflects the interests of both participants and sponsors in assessing and furthering research that can satisfy some urgent text processing needs in the very near term and can lead to solutions to more

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **MAY 1996**	2. REPORT TYPE	3. DATES COVERED **00-00-1996 to 00-00-1996**
4. TITLE AND SUBTITLE **Overview of Results of the MUC-6 Evaluation**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Command, Control, and Ocean Surveillance Center,RDT&E Division (NRaD),Information Access Technology Project Team, Code 44208,San Diego,CA,92152**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**		
13. SUPPLEMENTARY NOTES **TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996. Sponsored by the Defense Advanced Research Projects Agency.**		
14. ABSTRACT		
15. SUBJECT TERMS		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**20**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

challenging text understanding problems in the longer term. Identification of certain common types of names, which constitutes a large portion of the Named Entity task and a critical portion of the Template Element task, has proven to be largely a solved problem. Recognition of alternative ways of identifying an entity constitutes a large portion of the Coreference task and another critical portion of the Template Element task and has been shown to represent only a modest challenge when the referents are names or pronouns. The mix of challenges that the Scenario Template task represents has been shown to yield levels of performance that are smilar to those achieved in previous MUCs, but this time with a much shorter time required for porting.

Documentation of each of the tasks and summary scores for all systems evaluated can be found in the MUC-6 proceedings [1].

CORPUS

Testing was conducted using Wall Street Journal texts provided by the Linguistic Data Consortium. The articles used in the evaluation were drawn from a corpus of approximately 58,000 articles spanning the period of January 1993 through June 1994. This period comprised the "evaluation epoch." As a condition for participation in the evaluation, the sites agreed not to seek out and exploit Wall Street Journal articles from that epoch once the training phase of the evaluation had begun, i.e., once the scenario for the Scenario Template task had been disclosed to the participants.

The training set and test set each consisted of 100 articles and were drawn from the corpus using a text retrieval system called Managing Gigabytes, whose retrieval engine is based on a context-vector model, producing a ranked list of hits according to degree of match with a keyword search query. It can also be used to do unranked, Boolean retrievals. The Boolean retrieval method was used in the initial probing of the corpus to identify candidates for the Scenario Template task, because the Boolean retrieval is relatively fast, and the unranked results are easy to scan to get a feel for the variety of nonrelevant as well as relevant documents that match all or some of the query terms. Once the scenario had been identified, the ranked retrieval method was used, and the ranked list was sampled at different points to collect approximately 200 relevant and 200 nonrelevant articles, representing a variety of article types (feature articles, brief notices, editorials, etc.). From those candidate articles, the training and test sets were selected blindly, with later checks and corrections for imbalances in the relevant/nonrelevant categories and in article types.

From the 100 test articles, a subset of 30 articles (some relevant to the Scenario Template task, others not) was selected for use as the test set for the Named Entity and Coreference tasks. The selection was again done blindly, with later checks to ensure that the set was fairly representative in terms of article length and type. Note that although Named Entity, Coreference and Template Element are defined as domain-independent tasks, the articles that were used for MUC-6 testing were selected using domain-dependent criteria pertinent to the Scenario Template task. The manually filled templates were created with the aid of Tabula Rasa, a software tool developed for the Tipster Text Program by New Mexico State University Computing Research Laboratory.

NAMED ENTITY

The Named Entity (NE) task requires insertion of SGML tags into the text stream. The tag elements are ENAMEX (for entity names, comprising organizations, persons, and locations), TIMEX (for temporal expressions, namely direct mentions of dates and times), and NUMEX (for number expressions, consisting only of direct mentions of currency values and percentages). A TYPE attribute accompanies each tag element and identifies the subtype of each tagged string: for ENAMEX, the TYPE value can be ORGANIZATION, PERSON, or LOCATION; for TIMEX, the TYPE value can be DATE or TIME; and for NUMEX, the TYPE value can be MONEY or PERCENT.

Text strings that are to be annotated are termed *markables*. As indicated above, markables include names of organizations, persons, and locations, and direct mentions of dates, times, currency values and percentages. *Non-markables* include names of products and other miscellaneous names ("Macintosh," "Wall Street Journal" (in reference to the periodical as a physical object), "Dow Jones Industrial Average"); names of groups of people and miscellaneous usages of person names ("Republicans," "Gramm-Rudman," "Alzheimer['s]"); addresses and adjectival forms of location names ("53140 Gatchell Rd.," "American"); indirect and vague mentions of dates and times ("a few minutes after the hour," "thirty days before the end of the year"); and miscellaneous uses of numbers, including some that are similar to currency or percentage expressions ("[Fees] 1 3/4," "12 points," "1.5 times").

The evaluation metrics used for NE are essentially the same as those used for the two template-filling tasks, Template Element and Scenario Template. The following breakdowns of overall scores on NE are computed:

- by *slot*, i.e., for performance across tag elements, across TYPE attributes, and across tag strings;
- by *subcategorization*, i.e., for performance on each TYPE attribute separately;
- by *document section*, i.e., for performance on distinct subparts of the article, as identified by the SGML tags contained in the original text: <HL> ("headline"), <DD> ("document date"), <DATELINE>, and <TXT> (the body of the article).

NE Results Overall

Fifteen sites participated in the NE evaluation, including two that submitted two system configurations for testing and one that submitted four, for a total of 20 systems. As shown in table 1, performance on the NE task overall was over 90% on the F-measure for half of the systems tested, which includes systems from seven different sites. On the basis of the results of the dry run, in which two of the nine systems scored over 90%, we were not surprised to find official scores that were similarly high, but it was not expected that so many systems would enter the formal evaluation and perform so well.

It was also unexpected that one of the systems would match human performance on the task. Human performance was measured by comparing the 30 draft answer keys produced by the annotator at NRaD with those produced by the annotator at SAIC. This test measures the amount of variability between the annotators. When the outputs are scored in "key-to-response" mode, as though one annotator's output represented the "key" and the other the "response," the humans achieved an overall F-measure of 96.68 and a corresponding error per response fill (ERR) score of 6%. The top-scoring system, the baseline configuration of the SRA system, achieved an F-measure of 96.42 and a corresponding error score of 5%.

In considering the significance of these results from a general standpoint, the following facts about the test set need to be remembered:

F-Measure	Error	Recall	Precision
96.42	5	96	97
95.66	7	95	96
94.92	8	93	96
94.00	10	92	96
93.65	10	94	93
93.33	11	92	95
92.88	10	94	92
92.74	12	92	93
92.61	12	89	96
91.20	13	91	91
90.84	14	91	91
89.06	18	84	94
88.19	19	86	90
85.82	20	85	87
85.73	23	80	92
84.95	22	82	89

Table 1. Summary NE scores on primary metrics for the top 16 (out of 20) systems tested, in order of decreasing F-Measure (P&R) [1]

[1] Key to F-measure scores: BBN baseline configuration 93.65, BBN experimental configuration 92.88, Knight-Ridder 85.73, Lockheed-Martin 90.84, UManitoba 93.33, UMass 84.95, MITRE 91.2, NMSU CRL baseline configuration 85.82, NYU 88.19, USheffield 89.06, SRA baseline configuration 96.42, SRA "fast" configuration 95.66, SRA "fastest" configuration 92.61, SRA "nonames" configuration 94.92, SRI 94.0, Sterling Software 92.74.

- It represents just one style of writing (journalistic) and has a basic basic toward financial news and a specific bias toward the topic of the Scenario Template task.
- It was very small (only 30 articles). There were no markable time expressions in the test set, and there were only a few markable percentage expressions.

The results should also be qualified by saying that they reflect performance on data that makes accurate usage of upper and lower case distinctions. What would performance be on data where case provided no (reliable) clues and for languages where case doesn't distinguish names? SRA ran an experiment on an upper-case version of the test set that showed 85% recall and 89% precision overall, with identification of organization names presenting the greatest problem. That result represents nearly a 10-point decrease on the F-measure from their official baseline. The case-insensitive results would be slightly better if the task guidelines themselves didn't depend on case distinctions in certain situations, as when identifying the right boundary for the organization name span in a string such as "the Chrysler division" (currently, only "Chrysler" would be tagged).

NE Results on Some Aspects of Task

Figures 1 and 2 show the sample size for the various tag elements and TYPE values. Note that nearly 80% of the tags were ENAMEX and that almost half of those were subcategorized as organization names. As indicated in table 2, all systems performed better on identifying person names than on identifying organization or location names, and all but a few systems performed better on location names than on organization names. Organization names are varied in their form, consisting of proper nouns, general vocabulary, or a mixture of the two. They can also be quite long and complex and can even have internal punctuation such as a commas or an ampersand. Sometimes it is difficult to distinguish them from names of other types, especially from person names. Common organization names, first names of people, and location names can be handled by recourse to list lookup, although there are drawbacks: some names may be on more than one list, the lists will not be

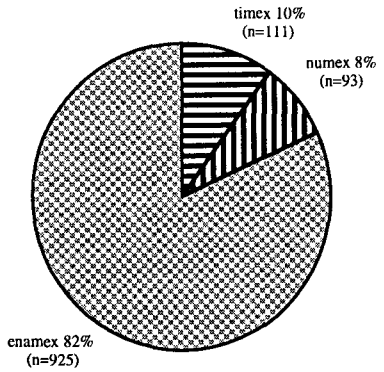

Figure 1. Distribution of NE tag elements in test set

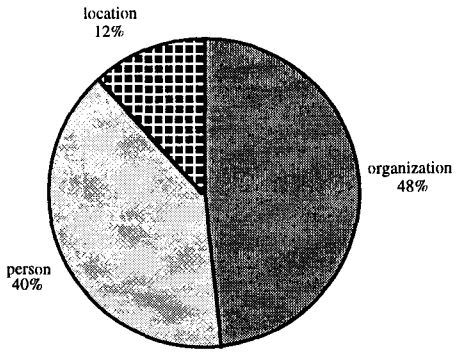

Figure 2. Subcategories of ENAMEX in test set

	enamex			timex		numex	
F-Measure	org	per	loc	date	time	money	percent
96.42	10	2	6	3	*	0	0
95.66	11	3	9	7	*	1	0
94.92	16	3	7	3	*	0	0
94.00	16	3	15	9	*	3	0
93.65	13	4	8	8	*	8	32
93.33	16	6	12	9	*	4	6
92.88	15	4	13	8	*	8	32
92.74	16	4	9	16	*	2	0
92.61	14	4	5	43	*	1	0
91.20	18	9	19	8	*	6	36
90.84	16	10	29	12	*	6	0
89.06	22	17	18	10	*	3	0
88.19	29	7	20	17	*	11	36
85.82	29	9	16	13	*	6	32
85.73	26	14	29	18	*	9	40
84.95	45	4	31	10	*	4	32

Table 2. NE subcategory scores (ERR metric), in order of decreasing overall F-Measure (P&R)

complete and may not match the name as it is realized in the text (e.g., may not cover the needed abbreviated form of an organization name, may not cover the complete person name), etc.

The difference that recourse to lists can make in performance is seen by comparing two runs made by SRA. The experimental configuration resulted in a three point decrease in recall and one point decrease in precision, compared to the performance of the baseline system configuration. The changes occurred only in performance on identifying organizations. BBN conducted a comparative test in which the experimental configuration used a larger lexicon than the baseline configuration, but the exact nature of the difference is not known and the performance differences are very small. As with the SRA experiment, the only differences in performance between the two BBN configurations are with the organization type. The University of Durham reported that they had intended to use gazetteer and company name lists, but didn't, because they found that the lists did not have much effect on their system's performance.

The error scores for persons, dates, and monetary expressions was less than or equal to 10% for the large majority of systems. Several systems posted scores under 10% error for locations, but none was able to do so for oganizations. For percentages, about half the systems had 0% error, which reflects the simplicity of that particular subtask. Note that the number of instances of percentages in the test set is so small that a single mistake could result in an error of 6%.

Slot-level performance on ENAMEX follows a different pattern for most systems from slot-level performance on NUMEX and TIMEX. The general pattern is for systems to have done better on the TEXT slot than on the TYPE slot for ENAMEX tags and for systems to have done better on the TYPE slot than on the TEXT slot for NUMEX and TIMEX tags. Errors on the TEXT slot are errors in finding the right span for the tagged string, and this can be a problem for all three subcategories of tag. The TYPE slot, however, is a more difficult slot for ENAMEX than for the other subcategories. It involves a three-way distinction for ENAMEX and only a two-way distinction for NUMEX and TIMEX, and it offers the possibility of confusing names of one type with names of another, especially the possibility of confusing organization names with person names.

Looking at the document section scores in table 3, we see that the error score on the body of the text was much lower than on the headline for all but a few systems. There was just one system that posted a higher error score on the body than on the headline, the baseline NMSU CRL configuration, and the difference in scores is largely due to the fact that the system overgenerated to a greater extent on the body than on the headline. Its basic strategy for

F-Measure	Document Date	Dateline	Headline	Text
96.42	0	0	8	5
95.66	0	0	7	7
94.92	0	0	8	8
94.00	0	0	20	9
93.65	0	2	16	10
93.33	0	4	38	9
92.88	0	0	18	10
92.74	0	0	22	11
92.61	100	0	18	9
91.20	0	0	30	13
90.84	3	11	19	14
89.06	3	4	28	18
88.19	0	0	22	20
85.82	0	6	18	21
85.73	0	44	53	21
84.95	0	0	50	21

Table 3. NE document subsection scores (ERR metric), in order of decreasing overall F-measure (P&R)

headlines was a conservative one: tag a string in the headline as a name only if the system had found it in the body of the text or if the system had predicted the name based on truncation of names found in the body of the text. Most, if not all, the systems that were evaluated on the NE task adopted the basic strategy of processing the headline after processing the body of the text.

The interannotator variability test provides reference points indicating human performance on the different aspects of the NE task. The document section results show 0% error on Document Date and Dateline, 7% error on Headline, and 6% error on Text. The subcategory error scores were 6% on Organization, 1% on Person, and 4% on Location, 8% on Date, and 0% on Money and Percent. These results show that human variability on this task patterns in a way that is similar to the performance of most of the systems in all respects except perhaps one: the greatest source of difficulty for the humans was on identifying dates. Analysis of the results shows that some Date errors were a result of simple oversight (e.g., "fiscal 1994") and others were a consequence of forgetting or misinterpreting the task guidelines with respect to determining the maximal span of the date expression (e.g., tagging "fiscal 1993's second quarter" and "Aug. 1" separately, rather than tagging "fiscal 1993's second quarter, ended Aug. 1" as a single expression in accordance with the task guidelines).

NE Results on "Walkthrough Article"

In the answer key for the walkthrough article there are 69 ENAMEX tags (including a few optional ones), six TIMEX tags and six NUMEX tags. Interannotator scoring showed that one annotator missed tagging one instance of "Coke" as an (optional) organization, and the other annotator missed one date expression ("September"). Common mistakes made by the systems included missing the date expression, "the 21st century," and spuriously identifying "60 pounds" (which appeared in the context, "Mr. Dooner, who recently lost 60 pounds over three-and-a-half months, ...") as a monetary value rather than ignoring it as a weight. In addition, a number of errors identifying entity names were made; some of those errors also showed up as errors on the Template Element task and are described in a later section of this paper.

COREFERENCE

The task as defined for MUC-6 was restricted to noun phrases (NPs) and was intended to be limited to phenomena that were relatively noncontroversial and easy to describe. The variety of high-frequency phenomena covered by the task is partially represented in the following hypothetical example, where all bracketed text segments are considered coreferential:

[Motor Vehicles International Corp.] announced a major management shake-up. ... [MVI] said the chief executive officer has resigned. ... [The Big 10 auto maker] is attempting to regain market share. ... [It] will announce significant losses for the fourth quarter. ... A [company] spokesman said [they] are moving [their] operations to Mexico in a cost-saving effort. ... [MVI, [the first company to announce such a move since the passage of the new international trade agreement],] is facing increasing demands from unionized workers. ... [Motor Vehicles International] is [the biggest American auto exporter to Latin America].

The example passage covers a broad spectrum of the phenomena included in the task. At one end of the spectrum are the proper names and aliases, which are inherently definite and whose referent may appear anywhere in the text. In the middle of the spectrum are definite descriptions and pronouns whose choice of referent is constrained by such factors as structural relations and discourse focus. On the periphery of the central phenomena are markables whose status as coreferring expressions is determined by syntax, such as predicate nominals ("Motor Vehicles International is the biggest American auto exporter to Latin America") and appositives ("MVI, the first company to announce such a move since the passage of the new international trade agreement"). At the far end of the spectrum are bare common nouns, such as the prenominal "company" in the example, whose status as a referring expression may be questionable.

An algorithm developed by the MITRE Corporation for MUC-6 was implemented by SAIC and used for scoring the task. The algorithm compares the equivalence classes defined by the coreference links in the manually-generated answer key and the system-generated response. The equivalence classes are the models of the identity equivalence coreference relation. Using a simple counting scheme, the algorithm obtains recall and precision scores by determining the minimal perturbations required to align the equivalence classes in the key and response. No metrics other than recall and precision were defined for this task, and no statistical significance testing was performed on the scores.

CO Results Overall

In all, seven sites participated in the MUC-6 coreference evaluation. Most systems achieved approximately the same levels of performance: five of the seven systems were in the 51%-63% recall

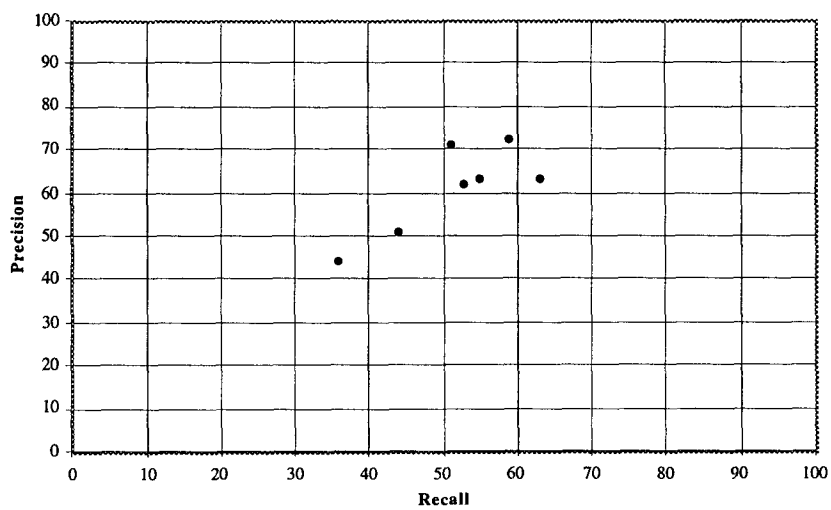

Figure 3. Overall recall and precision on the CO task[2]

[2] Key to recall and precision scores: UDurham 36R/44P, UManitoba 63R/63P, UMass 44R/51P, NYU 53R/62P, UPenn 55R/63P, USheffield 51R/71P, SRI 59R/72P.

range and 62%-72% precision range. About half the systems focused only on individual coreference, which has direct relevance to the other MUC-6 evaluation tasks.

A few of the evaluation sites reported that good name/alias recognition alone would buy a system a lot of recall and precision points on this task, perhaps about 30% recall (since proper names constituted a large minority of the annotations) and 90% precision. The precision figure is supported by evidence from the NE evaluation. In that evaluation, a number of systems scored over 90% on the named entity recall and precision metrics, providing a sound basis for good performance on the coreference task for individual entities.

In the middle of the effort of preparing the test data for the formal evaluation, an interannotator variability test was conducted. The two versions of the independently prepared, manual annotations of 17 articles were scored against each other using the scoring program in the normal "key to response" scoring mode. The amount of agreement between the two annotators was found to be 80% recall and 82% precision. There was a large number of factors that contributed to the 20% disagreement, including overlooking coreferential NPs, using different interpretations of vague portions of the guidelines, and making different subjective decisions when the text of an article was ambiguous, sloppy, etc. Most human errors pertained to definite descriptions and bare nominals, not to names and pronouns.

CO Results on Some Aspects of Task and on "Walkthrough Article"

To keep the annotation of the evaluation data fairly simple, the MUC-6 planning committee decided not to design the notation to subcategorize linkages and markables in any way. Two useful attributes for the equivalence class as a whole would be one to distinguish individual coreference from type coreference and one to identify the general semantic type of the class (organization, person, location, time, currency, etc.). For each NP in the equivalence class, it would be useful to identify its grammatical type (proper noun phrase, definite common noun phrase, bare singular common noun phrase, personal pronoun, etc.). The decision to minimize the annotation effort makes it difficult to do detailed quantitative analysis of the results.

An analysis by the participating sites of their system's performance on the walkthrough article provides some insight into performance on aspects of the coreference task that were dominant in that article. The article contains about 1000 words and approximately 130 coreference links, of which all but about a dozen are references to individual persons or individual organizations. Approximately 50 of the anaphors are personal pronouns, including reflexives and possessives, and 58 of the markables (anaphors and antecedents) are proper names, including aliases. The percentage of personal pronouns is relatively high (38%), compared to the test set overall (24%), as is the percentage of proper names (40% on this text versus an estimate of 30% overall).

Performance on this particular article for some systems was higher than performance on the test set overall, reaching as high as 77% recall and 79% precision. These scores indicate that pronoun resolution techniques as well as proper noun matching techniques are good, compared to the techniques required to determine references involving common noun phrases. For common noun phrases, the systems were not required to include the entire NP in the response; the response could minimally contain only the head noun. Despite this flexibility in the expected contents of the response, the systems nonetheless had to implicitly recognize the full NP, since to be considered coreferential, the head *and* its modifiers all had to be consistent with another markable.

TEMPLATE ELEMENT

The Template Element (TE) task requires extraction of certain general types of information about entities and merging of the information about any given entity before presentation in the form of a template (or "object"). For MUC-6 the entities that were to be extracted were limited to organizations and persons.[3] The ORGANIZATION object contains attributes ("slots") for the string representing the organization name (ORG_NAME), for strings representing any abbreviated versions of the name (ORG_ALIAS), for a string that describes the particular organization (ORG_DESCRIPTOR), for a subcategory of the type of organization (ORG_TYPE, whose permissible values are GOVERNMENT, COMPANY, and OTHER), and for canonical forms of the specific and general location of the organization (ORG_LOCALE and ORG_COUNTRY). The PERSON object contains

[3]The task documentation includes definition of an "artifact" entity, but that entity type was not used in MUC-6 for either the dry run or the formal run. The entity types that were involved in the evaluation are the same as those required for the Scenario Template task.

slots only for the string representing the person name (PER_NAME), for strings representing any abbreviated versions of the name (PER_ALIAS), and for strings representing a very limited range of titles (PER_TITLE).

The task places heavy emphasis on recognizing proper noun phrases, as in the NE task, since all slots except ORG_DESCRIPTOR and PER_TITLE expect proper names as slot fillers (in string or canonical form, depending on the slot. However, the organization portion of the TE task is not limited to recognizing the referential identity between full and shortened names; it requires the use of text analysis techniques at all levels of text structure to associate the descriptive and locative information with the appropriate entity. Analysis of complex NP structures, such as appositional structures and postposed modifier adjuncts, is needed in order to relate the locale and descriptor to the name in "Creative Artists Agency, the big Hollywood talent agency" and in "Creative Artists Agency, a big talent agency based in Hollywood." Analysis of sentence structures to identify grammatical relations such as predicate nominals is needed in order to relate those same pieces of information in "Creative Artists Agency is a big talent agency based in Hollywood." Analysis of discourse structure is needed in order to identify long-distance relationships.

The answer key for the TE task contains one object for each specific organization and person mentioned in the text. For generation of a PERSON object, the text must provide the name of the person (full name or part of a name). For generation of an ORGANIZATION object, the text must provide either the name (full or part) or a descriptor of the organization. Since the generation of these objects is independent of the relevance criteria imposed by the Scenario Template (ST) task, there are many more ORGANIZATION and PERSON objects in the TE key than in the ST key. For the formal evaluation, there were 606 ORGANIZATION and 496 PERSON objects in the TE key, versus 120 ORGANIZATION and 137 PERSON objects in the ST key.

The same set of articles was used for TE as for ST; therefore, the content of the articles is oriented toward the terms and subject matter covered by the ST task, which concerns changes in corporate management.[4] One effect of this bias is simply the number of entities mentioned in the articles: for the test set used for the MUC-6 dry run, which was based on a scenario concerning labor union contract negotiations, there were only about half as many organizations and persons mentioned as there were in the test set used for the formal run.

TE Results Overall

Twelve systems -- from eleven sites, including one that submitted two system configurations for testing -- were tested on the TE task. All but two of the systems posted F-measure scores in the 70-80% range, and four of the systems were able to achieve recall in the 70-80% range while maintaining precision in the 80-90% range, as shown in the figure 4. Human performance was measured in terms of variability between the outputs produced by the two NRaD and SAIC evaluators for 30 of the articles in the test set (the same 30 articles that were used for NE and CO testing). Using the scoring method in which one annotator's draft key serves as the "key" and the other annotator's draft key serves as the "response," the overall consistency score was 93.14 on the F-measure, with 93% recall and 93% precision.

TE Results on Some Aspects of Task

Given the more varied extraction requirements for the ORGANIZATION object, it is not surprising that performance on that portion of the TE task was not as good as on the PERSON object[5], as is clear in figure 5.

Figure 6 indicates the relative amount of error contributed by each of the slots in the ORGANIZATION object. It is evident that the more linguistic processing necessary to fill a slot, the harder the slot is to fill correctly. The ORG_COUNTRY slot is a special case in a way, since it is required to be filled when the ORG_LOCALE slot is filled. (The reverse is not the case, i.e., ORG_COUNTRY may be filled even if ORG_LOCALE is not, but this situation is relatively rare.) Since a missing or spurious ORG_LOCALE is likely to incur the same error in ORG_COUNTRY, the error scores for the two slots are understandably similar.

[4] The method used for selecting the articles for the test set is described at the beginning of this article.

[5] The highest score for the PERSON object, 95% recall and 95% precision, is close to the highest score on the NE subcategorization for person, which was 98% recall and 99% precision.

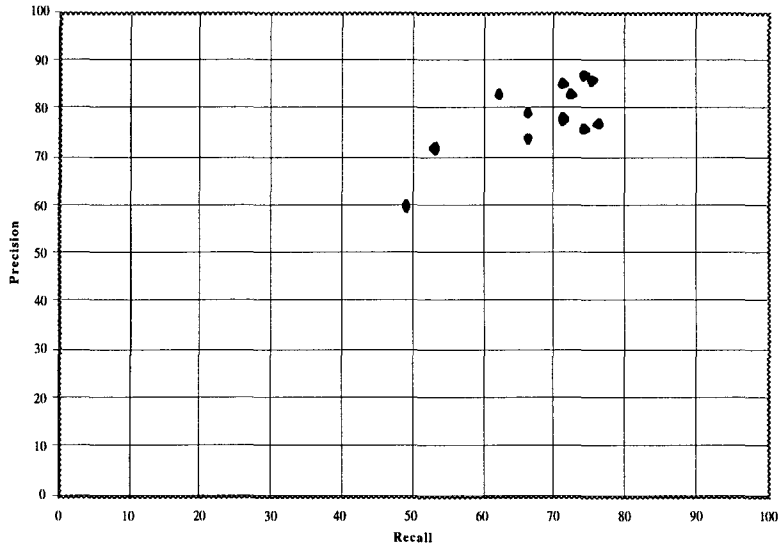

Figure 4. Overall recall and precision on the TE task[6]

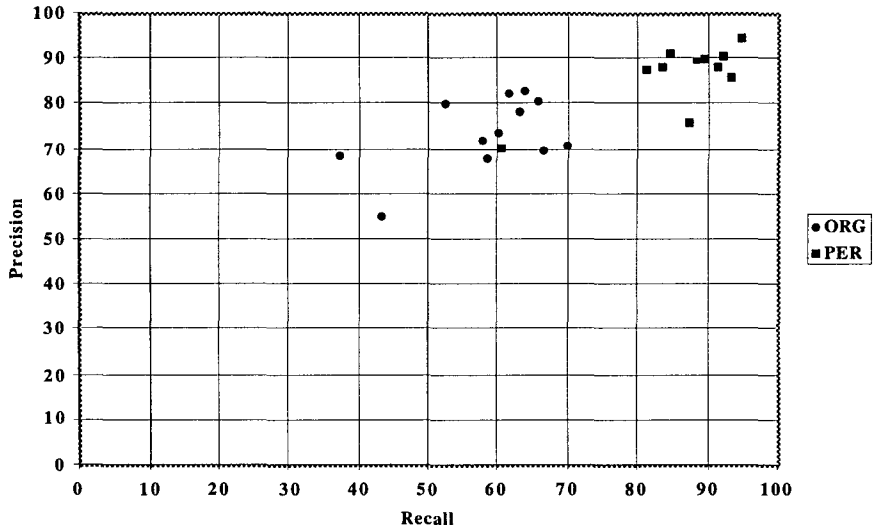

Figure 5. Organization and Person object recall and precision on the TE task

[6]Key to recall and precision scores: BBN 66R/79P, UDurham 49R/60P, Lockheed-Martin 76R/77P, UManitoba 71R/78P, UMass 53R/72P, MITRE 71R/85P, NYU 62R/83P, USheffield 66R/74P, SRA baseline configuration 75R/86P, SRA "noref" configuration 74R/87P, SRI 74R/76P, Sterling Software 72R/83P.

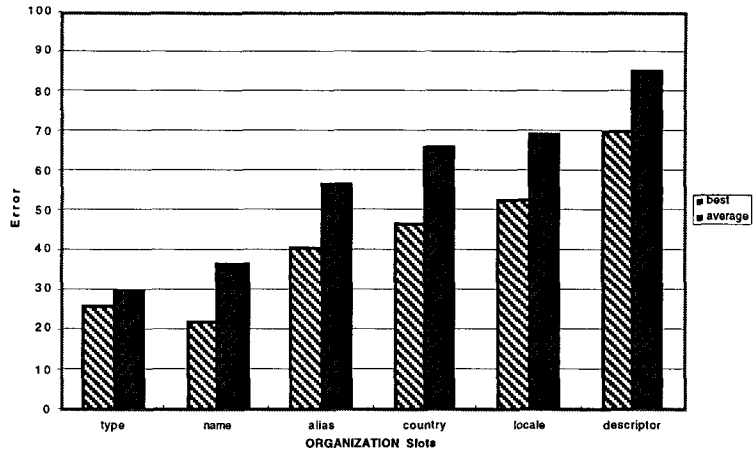

Figure 6. Best and average error per response fill Organization object slot scores for TE task

With respect to performance on ORG_DESCRIPTOR, note that there may be multiple descriptors (or none) in the text. However, the task does not require the system to extract all descriptors of an entity that are contained in the text; it requires only that the system extract one (or none). Frequently, at least one can be found in close proximity to an organization's name, e.g., as an appositive ("Creative Artists Agency, *the big Hollywood talent agency*"). Nonetheless, performance is much lower on this slot than on others.

Leaving aside the fact that descriptors are common noun phrases, which makes them less obvious candidates for extraction than proper noun phrases would be, what reasons can we find to account for the relatively low performance on the ORG_DESCRIPTOR slot? One reason for low performance is that an organization may be identified in a text solely by a descriptor, i.e., without a fill for the ORG_NAME slot and therefore without the usual local clues that the NP is in fact a relevant descriptor. It is, of course, also possible that a text may identify an organization solely by *name*. Both possibilities present increased opportunities for systems to undergenerate or overgenerate. Also, the descriptor is not always close to the name, and some discourse processing may be required in order to identify it -- this is likely to increase the opportunity for systems to miss the information. A third significant reason is that the response fill had to match the key fill exactly in order to be counted correct; there was no allowance made in the scoring software for assigning full or partial credit if the response fill only partially matched the key fill. It should be noted that human performance on this task was also relatively low, but it is unclear whether the degree of disagreement can be accounted for primarily by the reasons given above or whether the disagreement is attributable to the fact that the guidelines for that slot had not been finalized at the time when the annotators created their version of the keys.

TE Results on "Walkthrough Article"

TE performance of all systems on the walkthrough article was not as good as performance on the test set as a whole, but the difference is small for about half the systems. Viewed from the perspective of the TE task, the walkthrough article presents a number of interesting examples of entity type confusions that can result from insufficient processing. There are cases of organization names misidentified as person names, there is a case of a location name misidentified as an organization name, and there are cases of nonrelevant entity types (publications, products, indefinite references, etc.) misidentified as organizations. Errors of these kinds result in a penalty at the object level, since the extracted information is contained in the wrong type of object. Examples of each of these types of error appear below, along with the number of systems that committed the error. (An experimental configuration of the SRA system produced the same output as the baseline configuration and has been disregarded in the tallies; thus, the total number of systems tallied is eleven.)

1. Miscategorizations of entities as person (PER_NAME or PER_ALIAS) instead of organization (ORG_NAME or ORG_ALIAS)

- Six systems: *McCann-Erickson* (also extracted with the name of "McCann," "One McCann," "While McCann"; organization category is indicated clearly by context in which full name appears, "John Dooner Will Succeed James At Helm of McCann-Erickson" in headline and "Robert L. James, chairman and chief executive officer of McCann-Erickson, and John J. Dooner Jr., the agency's president and chief operating officer" in the body of the article)
- Six systems: *J. Walter Thompson* (also extracted with the name of "Walter Thompson"; organization category is indicated by context, "Peter Kim was hired from WPP Group's J. Walter Thompson last September...")
- Four systems: *Fallon McElligott* (organization category is indicated by context, "...other ad agencies, such as Fallon McElligott")
- One system: *Ammirati & Puris* (the presence of the ampersand is a clue, as is the context, "...president and chief executive officer of Ammirati & Puris"; but note that the article also mentions the name of one of the company's founders, Martin Puris)

2. Miscategorization of entity as organization (ORG_NAME) instead of location (ORG_LOCALE)
   - Two systems: *Hollywood* (location category is indicated by context, "Creative Artists Agency, the big Hollywood talent agency")
3. Miscategorization of nonrelevant entities as organization name, alias or descriptor (ORG_NAME, ORG_ALIAS, ORG_DESCRIPTOR)
   - Six systems: *New York Times* (publication name in phrase, "a framed page from the New York Times"; without sufficient context, the name can be ambiguous in its reference to a physical object versus an organization)
   - Three systems: *Coca-Cola Classic* (product name deriving from "Coca-Cola," which appears separately in several places in the article and is occasionally ambiguous even in context between product name and organization name)
   - One system: *Not Butter* (part of product name, "I Can't Believe It's Not Butter")
   - One system: *Taster* (part of product name, "Taster's Choice")
   - One system: *Choice* (part of product name, "Taster's Choice")
   - Five systems: *a hot agency* (nonspecific use of indefinite in phrase "...is interested in acquiring a hot agency")

Given the variety of contextual clues that must be taken into account in order to analyze the above entities correctly, it is understandable that just about any given system would commit at least one of them. But the problems are certainly tractable; none of the fifteen TE entities in the key (ten ORGANIZATION entities and five PERSON entities) was miscategorized by *all* of the systems.

In addition to miscategorization errors, the walkthrough text provides other interesting examples of system errors at the object level and the slot level, plus a number of examples of system successes. One success for the systems as a group is that each of the six smaller ORGANIZATION objects and four smaller PERSON objects (those with just one or two filled slots in the key) was matched perfectly by at least one system; in addition, one larger ORGANIZATION object and two larger PERSON objects were perfectly matched by at least one system. Thus, each of the five PERSON objects in the key and seven of the ten ORGANIZATION objects in the key were matched perfectly by at least one system. The three larger ORGANIZATION objects that none of the systems got perfectly correct are for the *McCann-Erickson*, *Creative Artists Agency*, and *Coca-Cola* companies. Common errors in these three ORGANIZATION objects included missing the descriptor or locale/country or failing to identify the organization's alias with its name.

SCENARIO TEMPLATE

A Scenario Template (ST) task captures domain- and task-specific information. Three scenarios were defined in the course of MUC-6: (1) a scenario concerning the event of organizations placing orders to buy aircraft with aircraft manufacturers (the "aircraft order" scenario); (2) a scenario concerning the event of contract negotiations between labor unions and companies (the "labor negotiations" scenario); (3) a scenario concerning changes in corporate managers occupying executive posts (the "management succession" scenario). The first scenario was used as an example of the general design of the ST task, the second was used for the MUC-6 dry run evaluation, and the third was used for the formal evauation. One of the innovations of MUC-6 was to formalize the general structure of event templates, and all three

scenarios defined in the course of MUC-6 conformed to that general structure. In this article, the management succession scenario will be used as the basis for discussion.

The management succession template consists of four object types, which are linked together via one-way pointers to form a hierarchical structure. At the top level is the TEMPLATE object, of which there is one instantiated for every document. This object points down to one or more SUCCESSION_EVENT objects if the document meets the event relevance criteria given in the task documentation. Each event object captures the changes occurring within a company with respect to one management post. The SUCCESSION_EVENT object points down to the IN_AND_OUT object, which in turn points down to PERSON Template Element objects that represent the persons involved in the succession event. The IN_AND_OUT object contains ST-specific information that relates the event with the persons. The ORGANIZATION Template Element objects are present at the lowest level along with the PERSON objects, and they are pointed to not only by the IN_AND_OUT object but also by the SUCCESSION_EVENT object. The organization pointed to by the event object is the organization where the relevant management post exists; the organization pointed to by the relational object is the organization that the person who is moving in or out of the post is coming from or going to.

The scenario is designed around the management post rather than around the succession act itself. Although the management post and information associated with it are represented in the SUCCESSION_EVENT object, that object does not actually represent an event, but rather a state, i.e., the vacancy of some management post. The relational-level IN_AND_OUT objects represent the personnel changes pertaining to that state.

ST Results Overall

Nine sites submitted a total of eleven systems for evaluation on the ST task. All the participating sites also submitted systems for evaluation on the TE and NE tasks. All but one of the development teams (UDurham) had members who were veterans of MUC-5.

Of the 100 texts in the test set, 54 were relevant to the management succession scenario, including six that were only marginally relevant. Marginally relevant event objects are marked in the answer key as being optional, which means that a system is not penalized if it does not produce such an event object. The approximate 50-50 split between relevant and nonrelevant texts was

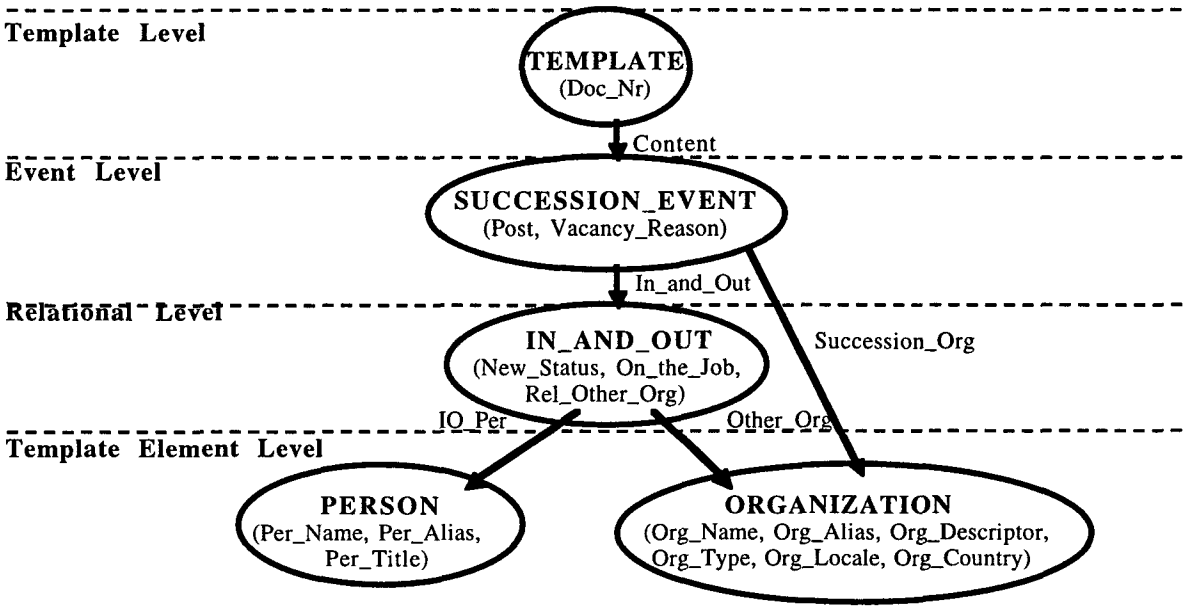

Figure 7. Management Succession Template Structure

intentional and is comparable to the richness of the MUC-3 "TST2" test set and the MUC-4 "TST4" test set. (The test sets used for MUC-5 had a much higher proportion of relevant texts.) Systems are measured for their performance on distinguishing relevant from nonrelevant texts via the *text filtering* metric, which uses the classic information retrieval definitions of recall and precision.

For MUC-6, text filtering scores were as high as 98% recall (with precision in the 80th percentile) or 96% precision (with recall in the 80th percentile). Similar tradeoffs and upper bounds on performance can be seen in the TST2 and TST4 results (see score reports in sections 2 and 4 of appendix G in [2]). However, performance of the systems as a group is better on the MUC-6 test set. The text filtering results for MUC-6, MUC-4 (TST4) and MUC-3 (TST2) are shown in figure 8.

Whereas the Text Filter row in the score report shows the system's ability to do text filtering (document detection), the All Objects row and the individual Slot rows show the system's ability to do information extraction. The measures used for information extraction include two overall ones, the F-measure and error per response fill, and several other, more diagnostic ones (recall, precision, undergeneration, overgeneration, and substitution). The text filtering definition of precision is different from the information extraction definition of precision; the latter definition includes an element in the formula that accounts for the number of spurious template fills generated.

The All Objects recall and precision scores are shown in figure 9. The highest ST F-measure score was 56.40 (47% recall, 70% precision). Statistically, large differences of up to 15 points may not be reflected as a difference in the ranking of the systems. Most of the systems fall into the same rank at the high end, and the evaluation does not clearly distinguish more than two ranks (see the paper on statistical significance testing by Chinchor in [1]). Human performance was measured in terms of interannotator variability on only 30 texts in the test set and showed agreement to be approximately 83%, when one annotator's templates were treated as the "key" and the other annotator's templates were treated as the "response."

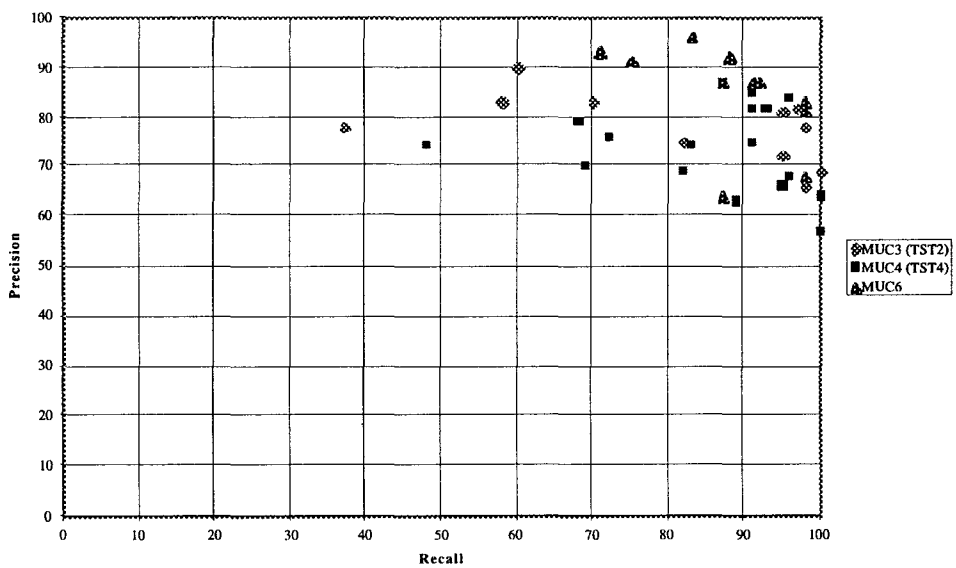

Figure 8. Text filtering recall and precision for scenario test sets with approximately 50% richness

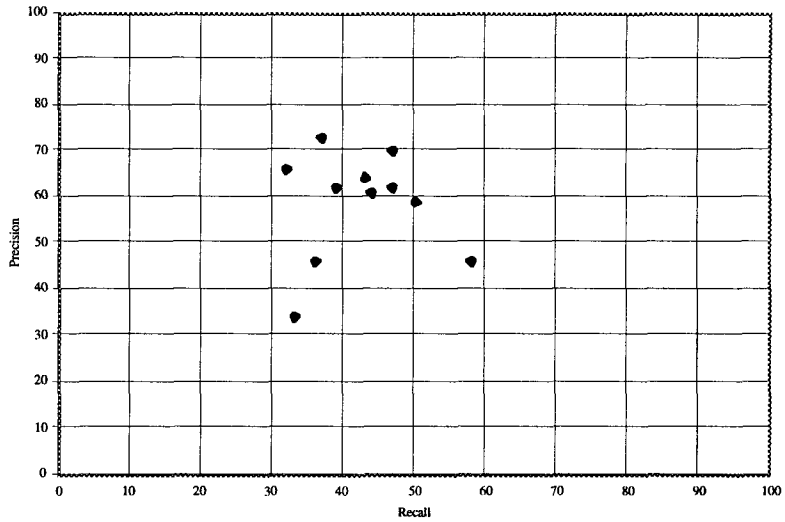

Figure 9. Overall information extraction recall and precision on the ST task[7]

No analysis has been done of the relative difficulty of the MUC-6 ST task compared to previous extraction evaluation tasks. The one-month limitation on development in preparation for MUC-6 would be difficult to factor into the computation, and even without that additional factor, the problem of coming up with a reasonable, objective way of measuring relative task difficulty has not been adequately addressed. Nonetheless, as one rough measure of progress in the area of information extraction as a whole, we can consider the F-measures of the top-scoring systems from the MUC-5 and MUC-6 evaluations.

MUC-6	56.40
MUC-5 EJV	52.75
MUC-5 JJV	60.07
MUC-5 EME	49.18
MUC-5 JME	56.31

Table 4. Highest P&R F-Measure scores posted for MUC-6 and MUC-5 ST tasks

Note that table 4 shows four top scores for MUC-5, one for each language-domain pair: English Joint Ventures (EJV), Japanese Joint Ventures (JJV), English Microelectronics (EME), and Japanese Microelectronics (JME). From this table, it may be reasonable to conclude that progress has been made, since the MUC-6 performance level is at least as high as for three of the four MUC-5 tasks and since that performance level was reached after a much shorter time.

ST Results on Some Aspects of Task and on "Walkthrough Article"

Three succession events are reported in the walkthrough article. Successful interpretation of three sentences from the walkthrough article is necessary for high performance on these events. The tipoff on the first two events comes at the end of the second paragraph:

> Yesterday, McCann made official what had been widely anticipated: Mr. James, 57 years old, is stepping down as chief executive officer on July 1 and will retire as chairman at the end of the year. He will be succeeded by Mr. Dooner, 45.

The basis of the third event comes halfway through the two-page article:

> In addition, Peter Kim was hired from WPP Group's J. Walter Thompson last September as vice chairman, chief strategy officer, world-wide.

[7]Key to recall and precision scores: BBN 50R/59P, UDurham 33R/34P, Lockheed-Martin 43R/64P, UManitoba 39R/62P, UMass 36R/46P, NYU 47R/70P, USheffield 37R/73P, SRA baseline configuration 47R/62P, SRA "precision" configuration 32R/66P, SRA "recall" configuration 58R/46P, SRI 44R/61P.

	Answer Key	*System Output*
Event #1	James out, Dooner in as CEO of McCann-Erickson as a result of James departing the workforce; James is still on the job as CEO; Dooner is not on the job as CEO yet, and his old job was with the same org as his new job.	James out, Dooner in as CEO of McCann-Erickson as a result of a reassignment of James; James is not on the job as CEO any more, and his new job is at the same as his old job; Dooner may or may not be on the job as CEO yet, and his old job was with the same org as his new job. (SRA *satie_base* system)
Event #2	James out, Dooner in as chairman of McCann-Erickson as a result of James departing the workforce; James is still on the job as chairman; Dooner is not on the job as chairman yet, and his old job was with the same org as his new job.	James out, Dooner in as chairman of McCann-Erickson as a result of James departing the workforce; James is not on the job as chairman any more; Dooner is already on the job as chairman, and his old job was with Ammirati & Puris. (NYU system)
Event #3	Kim in as "vice chairman, chief strategy officer, world-wide" of McCann-Erickson, where the vacancy existed for other/unknown reasons; he is already on the job in the post, and his old job was with J. Walter Thompson	Kim in as vice chairman of WPP Group, where the vacancy existed for other/unknown reasons; he may or may not be on the job in that post yet, and the article doesn't say where his old job was. (BBN system)

Table 5. Paraphrased summary of ST outputs for walkthrough article

The article was relatively straightforward for the annotators who prepared the answer key, and there were no substantive differences in the output produced by each of the two annotators.

Table 5 contains a paraphrased summary of the output that was to be generated for each of these events, along with a summary of the output that was actually generated by systems evaluated for MUC-6. The system-generated outputs are from three different systems, since no one system did better than all other systems on all three events. The substantive differences between the system-generated output and the answer key are indicated by underlining in the system output.

Recurring problems in the system outputs include the information about whether the person is currently on the job or not and the information on where the outgoing person's next job would be and where the incoming person's previous job was. Note also that even the best system on the third event was unable to determine that the succession event was occurring at McCann-Erickson; in addition, it only partially captured the full title of the post. To its credit, however, it did recognize that the event was relevant; only two systems produced output that is recognizable as pertaining to this event. One common problem was the simple failure to recognize "hire" as an indicator of a succession.

Two systems never filled the OTHER_ORG slot or its dependent slot, REL_OTHER_ORG, despite the fact that data to fill those slots was often present; over half the IN_AND_OUT objects in the answer key contain data for those two slots. Almost without exception, systems did more poorly on those two slots than on any others in the SUCCESSION_EVENT and IN_AND_OUT objects; the best scores posted were 70% error on OTHER_ORG (median score of 79%) and 72% error on REL_OTHER_ORG (median of 86%).

Performance on the VACANCY_REASON and ON_THE_JOB slots was better for nearly all systems. The lowest error scores were 56% on VACANCY_REASON (median of 70%) and 62% on ON_THE_JOB (median of 71%).

The slot that most systems performed best on is NEW_STATUS; the lowest error score posted on that slot is 47% (median of 55%). This slot has a limited number of fill options, and the right answer is almost always either IN or OUT, depending on whether the person involved is assuming a post (IN) or vacating a post (OUT). Performance on the POST slot was not quite as good; the lowest error was 52% (median of 65%). The POST slot requires a text string as fill, and there is no finite list of possible fills for the slot. As seen in the third event of the walkthrough article, the fill can be an extended title such as "vice chairman, chief strategy officer, world-wide." For most events, however, the fill is one of a large handful of possibilities, including "chairman," "president," "chief executive [officer]," "CEO," "chief operating officer," "chief financial officer," etc.

DISCUSSION: CRITIQUE OF TASKS

Named Entity

The primary subject for review in the NE evaluation is its limited scope. A variety of proper name types were excluded, e.g. product names. The range of numerical and temporal expressions covered by the task was also limited; one notable example is the restriction of temporal expressions to exclude "relative" time expressions such as "last week". Restriction of the corpus to Wall Street Journal articles resulted in a limited variety of markables and in reliance on capitalization to identify candidates for annotation.

Some work on expanding the scope of the NE task has been carried out in the context of a foreign-language NE evaluation conducted in the spring of 1996. This evaluation is called the MET (Multilingual Named Entity) and, like MUC-6, was carried out under the auspices of the Tipster Text program. The experience gained from that evaluation will serve as critical input to revising the Engish version of the task.

Coreference

Many aspects of the CO task are in definite need of review for reasons of either theory or practice. One set of issues concerns the range of syntactically governed coreference phenomena that are considered markable. For example, apposition as a markable phenomenon was restrictively defined to exclude constructs that could rather be analyzed as left modification, such as "chief executive Scott McNealy," which lacks the comma punctuation that would clearly identify "executive" as the head of an appositive construction. Another set of issues is semantic in nature and includes fundamental questions such as the validity of including type coreference in the task and the legitimacy of the implied definition of coreference versus reference. If an antecedent expression is nonreferential, can it nonetheless be considered coreferential with subsequent anaphoric expressions? Or can only referring expressions corefer? Finally, the current notation presents a set of issues, such as its inability to represent multiple antecedents, as in conjoined NPs, or alternate antecedents, as in the case of referential ambiguity.

In short, the preliminary nature of the task design is reflected in the somewhat unmotivated boundaries between markables and nonmarkables and in weaknesses in the notation. One indication of immaturity of the task definition (as well as an indication of the amount of genuine textual ambiguity) is the fact that over ten percent of the linkages in the answer key were marked as "optional." (Systems were not penalized if they failed to include such linkages in their output.) The task definition is now under review by a discourse working group formed in 1996 with representatives from both inside and outside the MUC commuity, including representatives from the spoken-language community.

Template Element

There are miscellaneous outstanding problems with the TE task. With respect to the ORGANIZATION and PERSON objects, there are issues such as rather fuzzy distinctions among the three organization subtypes and between the organization name and alias, the extremely limited scope of the person title slot, and the lack of a person descriptor slot. The ARTIFACT object, which was not used for either the dry run or the formal evaluation, needs to be reviewed with respect to its general utility, since its definition reflects primarily the requirements of the MUC-5 microelectronics task domain. There is a task-neutral DATE slot that is defined as a template element; it was used in the MUC-6 dry run as part of the labor negotiation scenario, but as currently defined, it fails to capture meaningfully some of the recurring kinds of date information. In particular, problems remain with normalizing various types of date expressions, including ones that are vague and/or require extensive use of calendar information.

Scenario Template

The issues with respect to the ST task relate primarily to the ambitiousness of the scenario templates defined for MUC-6. Although the management scenario contained only five domain-specific slots (disregarding slots containing pointers to other objects), it nonetheless reflected an interest in capturing as complete a representation of the basic event as possible. As a result, a few "peripheral" facts about the event were included that were difficult to define in the task documentation and/or were not reported clearly in many of the articles.

Two of the slots, VACANCY_REASON and ON_THE_JOB, had to be filled on the basis of inference from subtle linguistic cues in many cases. An entire appendix to the scenario definition is devoted to heuristics for filling the ON_THE_JOB slot. These two slots caused problems for the

annotators as well as for the systems. The annotators' problems with VACANCY_REASON may have had more to do with understanding what the scenario definition was saying than with understanding what the news articles were saying. The annotators' problems with ON_THE_JOB were probably more substantive, since the heuristics documented in the appendix were complex and sometimes hard to map onto the expressions found in the news articles. A third slot, REL_OTHER_ORG, required special inferencing on the basis of both linguistics and world knowledge in order to determine the corporate relationship between the organization a manager is leaving and the one the manager is going to. There may, in fact, be just one organization involved -- the person could be leaving a post at a company in order to take a different (or an additional) post at the same company.

Defining a generalized template structure and using Template Element objects as one layer in the structure reduced the amount of effort required for participants to move their system from one scenario to another. Further simplification may be advisable in order to focus on core information elements and exclude somewhat idiosyncratic ones such as the three slots described above. In the case of the management succession scenario, a proposal was made to eliminate the three slots discussed above and more, including the relational object itself, and to put the personnel information in the event object. Much less information about the event would be captured, but there would be a much stronger focus on the most essential information elements. This would possibly lead to significant improvements in performance on the basic event-related elements and to development of good end-user tools for incorporating some of the domain-specific patterns into a generic extraction system.

CONCLUSIONS

The results of the evaluation give clear evidence of the challenges that have been overcome and the ones that remain along dimensions of both breadth and depth in automated text analysis. The NE evaluation results serve mainly to document in the MUC context what was already strongly suspected:

1. Automated identification is extremely accurate when identification of lexical pattern types depends only on "shallow" information, such as the form of the string that satisfies the pattern and/or immediate context;
2. Automated identification is significantly less accurate when identification is clouded by uncertainty or ambiguity (as when case distinctions are not made, when organizations are named after persons, etc.) and must depend on one or more "deep" pieces of information (such as world knowledge, pragmatics, or inferences drawn from structural analysis at the sentential and suprasentential levels).

The vast majority of cases are simple ones; thus, some systems score extremely well -- well enough, in fact, to compete overall with human performance. Commercial systems are available already that include identification of those defined for this MUC-6 task, and since a number of systems performed very well for MUC-6, it is evident that high performance is probably within reach of any development site that devotes enough effort to the task. Any participant in a future MUC evaluation faces the challenge of providing a named entity identification capability that would score in the 90th percentile on the F-measure on a task such as the MUC-6 one.

The TE evaluation task makes explicit one aspect of extraction that is fundamental to a very broad range of higher-level extraction tasks. The identification of a name as that of an organization (hence, instantiation of an ORGANIZATION object) or as a person (PERSON object) is a named entity identification task. The association of shortened forms of the name with the full name depends on techniques that could be used for NE and CO as well as for TE. The real challenge of TE comes from associating other bits of information with the entity. For PERSON objects, this challenge is small, since the only additional bit of information required is the person's title ("Mr.," "Ms.," "Dr.," etc.), which appears immediately before the name/alias in the text. For ORGANIZATION objects, the challenge is greater, requiring extraction of location, description, and identification of the type of organization.

Performance on TE overall is as high as 80% on the F-measure, with performance on ORGANIZATION objects significantly lower (70th percentile) than on PERSON objects (90th percentile). Top performance on PERSON objects came close to human performance, while performance on ORGANIZATION objects fell significantly short of human performance, with the caveat that human performance was measured on only a portion of the test set. Some of the shortfall in performance on the ORGANIZATION object is due to inadequate discourse processing, which is needed in order to get some of the non-local instances of the ORG_DESCRIPTOR, ORG_LOCALE and ORG_COUNTRY slot fills.

In the case of ORG_DESCRIPTOR, the results of the CO evaluation seem to provide further evidence for the relative inadequacy of current techniques for relating entity descriptions with entity names.

Systems scored approximately 15-25 points lower (F-measure) on ST than on TE. As defined for MUC-6, the ST task presents a significant challenge in terms of system portability, in that the test procedure required that all domain-specific development be done in a period of one month. For past MUC evaluations, the formal run had been conducted using the same scenario as the dry run, and the task definition was released well before the dry run. Since the development time for the MUC-6 task was extremely short, it could be expected that the test would result in only modest performance levels. However, there were at least three factors that might lead one to expect higher levels of performance than seen in previous MUC evaluations:

1. The standardized template structure minimizes the amount of idiosyncratic programming required to produce the expected types of objects, links, and slot fills.
2. The fact that the domain-neutral Template Element evaluation was being conducted led to increased focus on getting the low-level information correct, which would carry over to the ST task, since approximately 25% of the expected information in the ST test set was contained in the low-level objects.
3. Many of the veteran participating sites had gotten to the point in their ongoing development where they had fast and efficient methods for updating their systems and monitoring their progress.

It appears that there is a wide variety of sources of error that impose limits on system effectiveness, whatever the techniques employed by the system. In addition, the short time frame allocated for domain-specific development naturally makes it very difficult for developers to do sufficient development to fill complex slots that either are not always expected to be filled or are not crucial elements in the template structure.

Sites have developed architectures that are at least as general-purpose techniques as ever, perhaps as a result of having to produce outputs for as many as four different tasks. Many of the sites have emphasized their pattern-matching techniques in discussing the strengths of their MUC-6 systems. However, we still have full-sentence parsing (e.g. USheffield, UDurham, UManitoba); we sometimes have expectations of "deep understanding" (cf. UDurham's use of a world model) and sometimes not (cf. UManitoba's production of ST output directly from dependency trees, with no semantic representation per se). Some systems completed all stages of analysis before producing outputs for any of the tasks, including NE. Six of the seven sites that participated in the coreference evaluation also participated in the MUC-6 information extraction evaluation, and five of the six made use of the results of the processing that produced their coreference output in the processing that produced their information extraction output.

The introduction of two new tasks into the MUC evaluations and the restructuring of information extraction into two separate tasks have infused new life into the evaluations. Other sources of excitement are the spinoff efforts that the NE and CO tasks have inspired that bring these tasks and their potential applications to the attention of new research groups and new customer groups. In addition, there are plans to put evaluations on line, with public access, starting with the NE evaluation; this is intended to make the NE task familiar to new sites and to give them a convenient and low-pressure way to try their hand at following a standardized test procedure. Finally, a change in administration of the MUC evaluations is occurring that will bring fresh ideas. The author is turning over government leadership of the MUC work to Elaine Marsh at the Naval Research Laboratory in Washington, D.C. Ms. Marsh has many years of experience in computational linguistics to offer, along with extensive familiarity with the MUC evaluations, and will undoubtedly lead the work exceptionally well.

ACKNOWLEDGEMENTS

The definition and implementation of the evaluations reported on at the Message Understanding Conference was once again a "community" effort, requiring active involvement on the part of the evaluation participants as well as the organizers and sponsors. Individual thanks go to Ralph Grishman of NYU for serving as program co-chair, to Nancy Chinchor for her critical efforts on virtually all aspects of MUC-6, and to the other members of the program committee, which included Chinatsu Aone of SRA Corp., Lois Childs of Lockheed Martin Corp., Jerry Hobbs of SRI International, Boyan Onyshkevych of the U.S. Dept. of Defense, Marc Vilain of The MITRE Corp., Takahiro Wakao of the Univ. of Sheffield, and Ralph Weischedel of BBN Systems and Technologies. The author would also like to

acknowledge the critical behind-the-scenes computer support rendered at NRaD by Tim Wadsworth, who passed away suddenly in August 1995, leaving a lasting empty spot in my work and my heart.

REFERENCES

[1] *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, November 1995, San Francisco: Morgan Kaufmann.

[2] *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, June 1992, San Mateo: Morgan Kaufmann.