THE MESSAGE UNDERSTANDING CONFERENCES

Beth M. Sundheim

Naval Command, Control, and Ocean Surveillance Center RDT&E Division (NRaD)
53140 Gatchell Road, San Diego, CA 92152-7420
sundheim@nosc.mil

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, co-chaired by Ralph Grishman (NYU) and Beth Sundheim (NRaD). 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 previous MUC evaluations. All except the Scenario Template task are defined independently of any particular domain.

The evolution and design of the MUC-6 evaluation are described in the conference proceedings [1]. A basic characterization of the challenge presented by each 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.

Testing was conducted using Wall Street Journal texts provided by the Linguistic Data Consortium. The test set for the two information extraction tasks consisted of 100 articles. A subset of 30 articles was selected for use as the test set for the two SGML annotation tasks. The evaluation began with the distribution of the scenario definition and training data at the beginning of September. The test data was distributed four weeks later, with results due by the end of the week. Sixteen sites participated in the evaluation; 15 systems were evaluated for the NE task, 7 for CO, 11 for TE, and 9 for ST.[1]

The variety of tasks that were 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 challenging text understanding problems in the longer term. The hard work carried out by the planning committee over nearly two years led to extremely interesting and useful evaluation results.

- Identification of names, which constitutes a large portion of the NE task and a critical portion of the TE task, has proven to be largely a solved problem. The majority of systems evaluated on NE had recall and precision over 90%; the highest-scoring system had a recall of 96% and a precision of 97%, which was judged to be comparable to human performance on the task.

- Recognition of alternative ways of identifying an entity constitutes a large portion of the CO task and another critical portion of the TE task; it has been shown to represent only a modest challenge when the referents are names or pronouns. All but two of the TE systems posted combined recall-precision (F-measure) scores in the 70-80% range; four of the systems were able to achieve recall in the 70-80% range while maintaining precision in the 80-90% range. The top-scoring system had 75% recall, 86% precision. Five of the seven CO systems were in the 51%-63% recall range and 62%-72% precision range.

---

[1] The participating sites were BBN Systems and Technology, University of Durham (UK), Knight-Ridder Information, Lockheed-Martin, University of Manitoba (Canada), University of Massachusetts (Amherst), The MITRE Corp., New Mexico State University Computing Research Laboratory, New York University, University of Pennsylvania, SAIC (McLean), University of Sheffield (UK), Systems Research and Applications, SRI International, Sterling Software, and Wayne State University.

•The ST task concerned changes in corporate executive management personnel; the extracted information includes answers to the basic questions of "Who is creating or filling what vacancy at what organization?". The mix of challenges that the task represents -- extraction of domain-specific events and relations along with the pertinent entities (template elements) -- yielded levels of performance that are similar to those achieved in previous MUCs (40%-50% recall, 60%-70% precision), but with a much shorter time required for porting. The highest ST performance overall was 47% recall and 70% precision.

| 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)
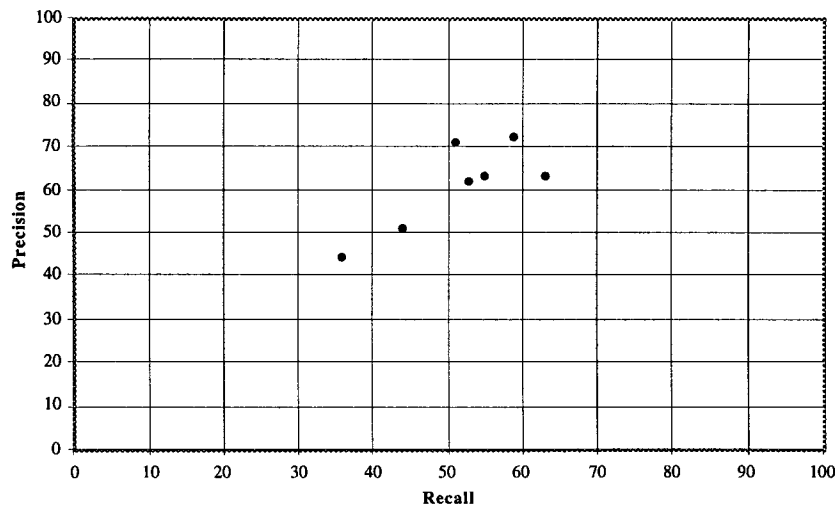
Figure 1. Overall recall and precision on the CO task
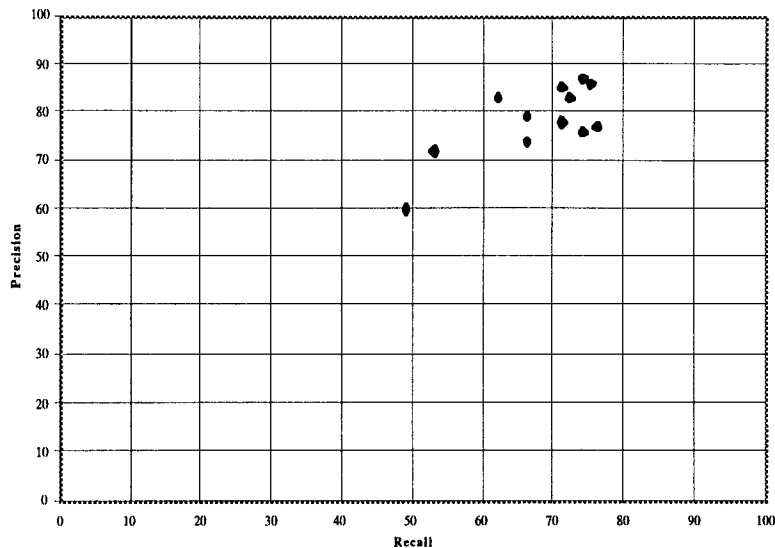
Figure 2. Overall recall and precision on the TE task
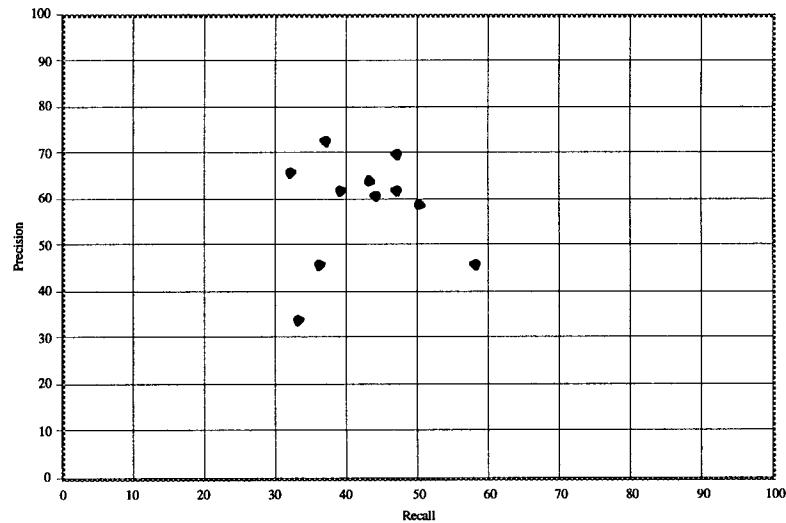
Figure 3. Overall information extraction recall and precision on the ST task

MUC-7 will be held in 1997, with Government coordination led by Elaine Marsh of the Naval Research Laboratory. Ms. Marsh is currently Section Head for the Intelligent Multimodal Multimedia (IM4) Section at the Navy Center for Artificial Intelligence. There she has conducted basic and exploratory research in natural language understanding and multimodal interactive systems. Prior to joining the Naval Research Laboratory, Ms. Marsh was employed as a research scientist on the Linguistic String Project at New York University.

She holds M.A. degrees from the University of Wisconsin-Madison and New York University and has completed additional graduate courses at New York University.

REFERENCES

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

37