# TIPSTER Information Extraction Evaluation: The MUC-7 Workshop

*Elaine Marsh*
Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
4555 Overlook Ave., SW
Washington, DC 20375-5337
Email: marsh@aic.nrl.navy.mil

## INTRODUCTION

The last of the "Message Understanding Conferences", which were designed to evaluate text extraction systems, was held in April 1998 in Fairfax, Virginia. The workshop was co-chaired by Elaine Marsh and Ralph Grishman. A group of 18 organizations, both from the United States and abroad, participated in the evaluation. MUC-7 introduced a wider set of tasks with larger sets of training and formal data than previous MUCs. Results showed that while performance on the named entity and template elements task remains relatively high, additional research is still necessary for improved performance on more difficult tasks such as coreference resolution and domain-specific template generation from textual sources.

## EVALUATION TASKS

MUC-7 consisted of six information extraction tasks. The Named Entity Task [NE] required systems to 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. The guidelines were brought in line with the multilingual task. The Multi-lingual Entity Task [MET] involved the execution of the NE task for Chinese and Japanese language texts. The Template Element Task [TE] required participants to extract basic information related to organization, person, and artifact entities, drawing evidence from anywhere in the text. The Template Relation Task [TR} was a new task for MUC-7, involving extracting relational information on generic domain-independent relations such as employee_of, manufacturer_of, and location_of relations. The Scenario Template Task [ST] consisted of extracting prespecified event information and relating that event information to particular organization, person, or artifact entities involved in that event. The final task was the Coreference Task, which involved capturing information on coreferring expressions, i.e. all mentions of a given entity, including those tagged in NE and TE tasks.

18 sites participated in MUC-7. 8 of 18 were university research groups. 8 were from outside the United States. Sites could participate in one or more of the tasks. 12 sites participated in the NE task, 9 in TE, 5 in TR, 5 in TE, and 5 in CO. No site participated in all of the tasks.

## TRAINING AND DATA SETS

The corpus for MUC-7 consisted of subsets of articles selected from a set of approximately 158,000 articles from the New York Times News Service (supplied by the LDC). The evaluation epoch of the articles was January 1- September 11, 1996. Training and test sets were retrieved from the corpus using the Managing Gigabytes text retrieval system using domain relevant terms. 2 sets of 100 articles from the aircraft accident domain were used for preliminary training, including the dryrun. 2 sets of 100 articles from the launch event domain were selected for the formal run after having been balanced for relevancy, type and source.

The training data set consisted of training keys for NE, TE, and TR tasks made available from a preliminary set of 100 articles; CO from a preliminary training set of 30 articles. A formal training set of 100 articles and answer keys were provided for the ST task.

The test set for the evaluation consisted of 100 articles and answer keys for NE (from the Formal Training data set) and 100 articles and answer keys for TE, TR, and ST. A subset of 30 articles and answer keys were provided for the CO task.

## FORMAL EVALUATION

The evaluation began with the distribution of the formal run test for NE at the beginning of March 1998. The training set of articles, ST guidelines and keys were made available at the beginning of March and one month afterward the test set of articles was made available by electronic transfer from SAIC. The deadline for completing the TE, TR, ST, and CO tasks was 6 April 1998 via electronic file transfer of system outputs to SAIC.

Tests were run by individual participating sites at their own facilities, following a written test procedure. Sites could conduct official "optional" tests in addition to the basic test and adaptive systems were permitted. Each site's system output was scored according to the following categories with respect to the answer keys: correct, incorrect, missing, spurious, possible (affected by inclusion and omission of optional data) and actual. Metrics included recall (a measure of how much of the key's fills were produced in the response), precision (a measure of how much of the response fills are actually in the key), F-measure (combining recall and precision into one measure, and ERR (error per response fill). Additional supporting metrics of undergeneration, overgeneration, and substitution were provided as well. The scoring procedure was completely automatic. Initial results for five tasks are presented in Figure 1.

In MUC-7, the new Template Relation task was an attempt to move from identifying domain-independent elements to identifying domain-independent relations that hold between these elements. The hope was that this would lead to performance improvements on the Scenario Template task. The evaluation domain for MUC-7 was concerned with vehicle launch events. The template consisted of one high-level event object with 7 slots, including two relational objects, three set fills, and two pointers to low-level objects. The domain represented a change from person-oriented domain of MUC-6 to a more artifact-oriented domain.

While there have been important advances in information extraction for named entity tasks and substantial improvement in the other tasks for which these MUC evaluations were developed, much remains to be done to put production-level information extraction systems on users' desks. We leave these breakthroughs to future researchers with thanks and recognition of the groundbreaking efforts of all the MUC participants throughout the years.

## ACKNOWLEDGMENTS

The MUC-7 evaluation was organized by NRL with technical support from SAIC and financial support from the DARPA TIPSTER Program. Other evaluation support was provided by the MUC-7 Program Committee, DARPA/ITO and the Tipster Executive Committee. The Linguistic Data Consortium provided all textual data for data selection and also provided the training and test data to participating sites at minimal cost.

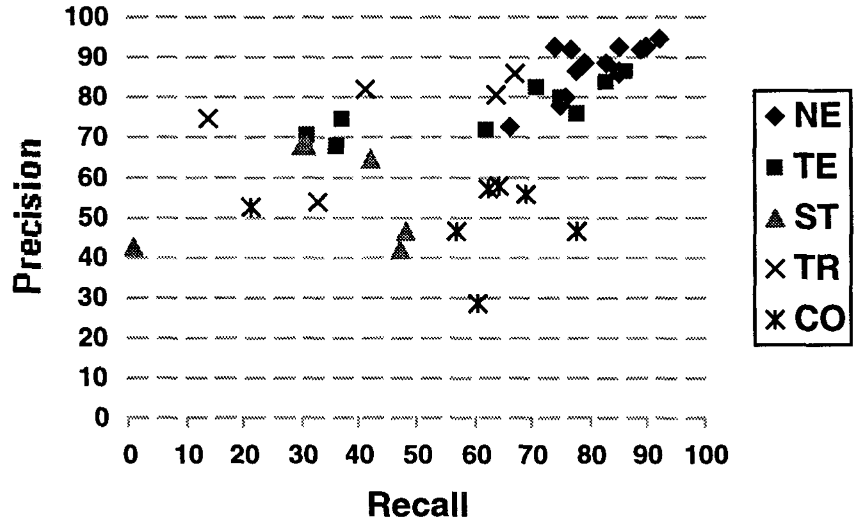

**Figure 1:** Overall recall and precision on all tasks