OVERVIEW OF MUC-7/MET-2

Nancy A. Chinchor
Science Applications International Corporation
10260 Campus Pt. Dr.
San Diego, CA 92121
chinchor@gso.saic.com

Overviews of English and Multilingual Tasks

The tasks performed by the systems participating in the seventh Message Understanding Conference and the Second Multilingual Entity Task are described here in general terms with examples.

Entities

On the level of entity extraction, Named Entities (NE) were defined as proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts. The annotation was SGML within the text stream. An example from MUC-7 (New York Times News Service) in English follows.

> The <ENAMEX TYPE="LOCATION">U.K.</ENAMEX> satellite television broadcaster said its subscriber base grew <NUMEX TYPE="PERCENT">17.5 percent</NUMEX> during <TIMEX TYPE="DATE">the past year</TIMEX> to 5.35 million

The Named Entity task was carried out in Chinese and Japanese (MET-2) concurrently with English (MUC-7).

Equivalence Classes

The task of Coreference (CO) had its origins in Semeval, an attempt after MUC-5 to define semantic research tasks that needed to be solved to be successful at generating scenario templates. In the MUC evaluations, only coreference of type identity was marked and scored [3]. The following example from MUC-7 (New York Times News Service) illustrates identity coreference between "its" and "The U.K. satellite television broadcaster" as well as that between the function "its subscriber base" and the value "5.35 million."

> *The U.K. satellite television broadcaster* said **its* subscriber base* grew 17.5 percent during the past year to *5.35 million*

The coreference task is a bridge between the NE task and the TE task.

Attributes

The attributes of entities are slot fills in Template Elements (TE) that consist of name, type, descriptor, and category slots. The attributes in the Template Element serve to further identify the entity beyond the name level.

All aliases are put in the NAME slot. Persons, organizations, artifacts, and locations are all TYPEs of Template Elements. All substantial descriptors used in the text appear in the DESCRIPTOR slot. The CATEGORY slot contains categories dependent on the element involved: persons can be civilian, military, or other; organizations can be government, company, or other; artifacts are limited to vehicles and can be for traveling on land, water, or in air; locations can be city, province, country, region, body of water, airport, or unknown. An example of a Template Element from MUC-7 follows:

> <ENTITY-9602040136-11> :=
> ENT_NAME: "Dennis Gillespie"
> ENT_TYPE: PERSON
> ENT_DESCRIPTOR: "Capt."
> / "the commander of Carrier Air Wing 11"

ENT_CATEGORY: PER_MIL

Facts

The Template Relations (TR) task marks relationships between template elements and can be thought of as a task in which well-defined facts are extracted from newswire text.

In MUC-7, we limited TR to relationships with organizations: employee_of, product_of, location_of. However, the task is easily expandable to all logical combinations and relations between entity typesAn example of Template Relations from MUC-7 follows:

<EMPLOYEE_OF-9602040136-5> :=
PERSON: <ENTITY-9602040136-11>
ORGANIZATION: <ENTITY-9602040136-1>

<ENTITY-9602040136-11> :=
ENT_NAME: “Dennis Gillespie”
ENT_TYPE: PERSON
ENT_DESCRIPTOR: “Capt.”
/ “the commander of Carrier Air Wing 11”
ENT_CATEGORY: PER_MIL

<ENTITY-9602040136-1> :=
ENT_NAME: “NAVY”
ENT_TYPE: ORGANIZATION
ENT_CATEGORY: ORG_GOVT

Events

The Scenario Template (ST) was built around an event in which entities participated. The scenario provided the domain of the dataset and allowed for relevancy judgments of high accuracy by systems.

The task definition for ST required relevancy and fill rules. The choice of the domain was dependent to some extent on the evaluation epoch. The structure of the template and the task definition tended to be dependent on the author of the task, but the richness of the templates also served to illustrated the utility of information extraction to users most effectively.

The filling of the slots in the scenario template was generally a difficult task for systems and a relatively large effort was required to produce ground truth. Reasonable agreement(>80%) between annotators was possible, but required sometimes ornate refinement of the task definition based on the data encountered.

How MUC-7 Differed from Previous MUCs

For the first time, the multilingual NE evaluation was run using training and test articles from comparable domains for all languages. The domain for all languages for training was airline crashes and the domain for all languages for testing was launch events. The domain change between the dry run and the formal run caused similar effects across languages. Sites expressed disappointment in their formal test scores when compared with their development test scores, but the formal test scores were still above the 80% operational threshold set by customers without any changes being made to systems for the domain change.

In MUC-7, there were more international sites participating than ever before. The papers reflect interesting observations by system developers who were non-native speakers of the language of their system and system developers who were native speakers of the language of their system.

In MUC-7, more data was provided for training and dry run and it was maintained through all of the updates to the guidelines during the evaluation cycle. The markup will be publicly available on the MUC website at http:\\www.muc.saic.com in the form of offsets from the beginning of each document. The rights to the documents

themselves can be purchased from the Linguistic Data Consortium (LDC).

The task definitions for MUC-7 were improved by having authors other than the original authors revise each of the guidelines for internal consistency and to dovetail into the other tasks evaluated. The communal effort in polishing the guidelines and the data markup noticeably improved the evaluation..

Table 1: Tasks Evaluated in MUC-3 through MUC-7

Evaluation\Tasks	Named Entity	Coreference	Template Element	Template Relation	Scenario Template	Multilingual
MUC-3					YES	
MUC-4					YES	
MUC-5					YES	YES
MUC-6	YES	YES	YES		YES	
MUC-7	YES	YES	YES	YES	YES	
MET-1	YES					YES
MET-2	YES					YES

Table 2: Maximum Results Reported in MUC-3 through MUC-7 by Task

Evaluation\Tasks	Named Entity	Coreference	Template Element	Template Relation	Scenario Template	Multilingual
MUC-3					R < 50% P < 70%	
MUC-4					F < 56%	
MUC-5					EJV F < 53% EME F < 50%	JJV F < 64% JME F < 57%
MUC-6	F < 97%	R < 63% P < 72%	F < 80%		F < 57%	
MUC-7	F < 94%	F < 62%	F < 87%	F < 76%	F < 51%	
Multilingual						
MET-1	C F < 85% J F < 93% S F < 94%					
MET-2	C F < 91% J F < 87%					

Legend: R = Recall P = Precision F = F-Measure with Recall and Precision Weighted Equally
E = English C = Chinese J = Japanese S = Spanish
JV = Joint Venture ME = Microelectronics

Brief History of the Message Understanding Conferences

MUC-7 is the last in the series of Message Understanding Conference Evaluations. So it is fitting to give a brief history of the Evaluations that were publicly funded. The major parameters of growth over the years were the tasks and the results. The tables below review these changes beginning with MUC-3. The two earlier evaluations were initiated, designed, and carried out by Beth Sundheim under the auspices of the Navy and focused on extraction from

military messages. Those evaluations listed here have been carried out under the auspices of the Tipster Text Program and focused on extraction from newswire articles. Scoring automation and other tools were supported under this program as well as research in evaluation methodology.

Guide to the MUC-7 Proceedings

The Table of Contents shows participants by task and language. Each site submitted paper(s) covering their task(s). For a separate list of which sites performed each of the tasks described above, please refer to Elaine Marsh's slides in this volume.

The appendices to the proceedings contain test materials and other supporting materials that augment the papers in the proceedings. For each of the tasks, a walkthrough article was chosen to allow all of the sites participating in that task to discuss their system response for a common article. The walkthrough articles and the answer keys for each task appear first.

Following the walkthroughs are the formal task definitions provided to all of the sites participating. The datasets discussed were all marked up by human annotators following these guidelines. Next are the score reports output by the automatic scorer which compared the system responses on each task to the human generated answer keys for the formal run test articles. The statistical results represent the significance groupings of the sites for each task based on an approximate randomization algorithm run on the document-by-document scores for each pair of sites. For Named Entity in English, the human annotators' scores are given and included in the statistical significance testing because the systems can achieve scores that are close to human performance. The annotators were significantly better than the systems. Finally, there is the User's Manual for the automated scorer which is in the public domain.

Acknowledgments

We would like to acknowledge DARPA as our major funding agency throughout our work on the evaluations. The Tipster Text Program and especially the Tipster Executive Committee contributed greatly to the success of the evaluations and the development of extraction technologies. We would like to thank especially Beth Sundheim for her vision in building evaluations which both fostered and tested the development of text extraction over the last decade. These evaluations were a community effort of sponsors, Program Committees, and all of the participants over the years. We would also like to acknowledge the contribution of the Linguistic Data Consortium in providing newswire data for the evaluations following MUC-4 and the contribution of Morgan Kaufman Publishers for publishing the Proceedings of the Message Understanding Conferences.