MUC/MET Evaluation Trends

Nancy A. Chinchor
Science Applications International Corporation
10260 Campus Point Drive
San Diego, CA 92121
chinchor@gso.saic.com
(619)458-2614

INTRODUCTION

During the course of the Tipster Program, evaluation methodology for information extraction developed as the technology progressed. Multiple task levels and multiple languages were successful targets of information extraction. Automated scoring and statistical significance algorithms were developed for use in scoring systems and for interannotator agreement measures. The scoring interface allowed both system developers and annotators to analyze errors and improve their work. This software and the marked datasets are now in the public domain. Future projects are being carried out based on simplifications indicated by the data, downstream applications, and tractability of scoring algorithms.

EVALUATION METHODOLOGY

Tasks

The original MUC task was to extract information about relevant events from newswire texts and use it to fill the slots in a scenario template. In MUC-3 and MUC-4 the terrorism scenario was used for the entire evaluation cycle including training, dry run, and formal run [1].

In MUC-5, the two domains of joint ventures and microelectronics fabrication capabilities were used in English and Japanese for the entire development cycle.

In MUC-6, a transition to focus on portability limited the time that developers could know the domain of the scenario template for the test so training and test data were provided for labor negotiations in the dry run and for management succession in the formal run, each lasting one month. Also, in MUC-6, the task was broken down into markup of named entities and coreference as well as template elements which were then used in the scenario template.

After MUC-6, a dry run of MET, the multilingual extraction task, was done for named entity in Chinese, Japanese, and Spanish [2]. A full evaluation of named entity markup in Chinese and Japanese was done in MET-2 as part of MUC-7.

The template relation task was also added in MUC-7 to allow the extraction of relationships between template elements unrelated to scenario template. The domain for the MUC-7 dry run of scenario template was

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

air crashes and the formal run was launch events.

The table above summarizes this development of tasks over the Tipster years of MUC/MET. The tasks are defined in more detail in the remainder of this section. The results of the evaluations are given in the next section in tabular format.

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) 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

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 types

An 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.

Task Definitions

As experience was gained in defining tasks for information extraction, certain principles became invaluable. It was important for the utility of the task to be apparent to end users. The lower level tasks needed to dovetail into the higher level tasks requiring relatively more processing for each higher level.

It was important for the task definitions to allow the achievement of an 80 - 99% threshold in interannotator agreement depending on how well systems were performing. Also, the ability to annotate text rapidly and with ease was critical to the end product: guidelines and datasets of high quality for research and development.

The process of refining the task definition required concentration on several cycles of independent annotation, analysis of annotator agreement, and detailed note-taking on which examples required updates to the task definition. The production of consistent datasets was always the goal.

Datasets

To be sure that systems could work on newswire from different sources, the different evaluations utilized material from various sources. MUC-3 and MUC-4 used articles from the Foreign Broadcast Information Service (FBIS). MUC-5 and MUC-6 used the well-edited Wall Street Journal (WSJ). MUC-7 used articles from the New York Times News Service (NYTNS) which contained journalism from multiple news organizations in a uniform SGML format.

Typically, there were 100 texts per dataset. The texts were chosen using mixtures of keywords associated with the domains (terrorism, joint ventures, microelectronics, labor relations, management succession, air crashes, launch events). A pre-defined relevancy ratio was used for each dataset, usually 65% of the texts were relevant. Datasets in MUC-7 were provided for general training, dry run training and test, and formal run training and test data.

In all of the MUC/MET evaluations the annotation accuracy was at least 80%, or higher whenever system performance was closer to human performance. We coordinated adjudication after the evaluation results were reported but before the final package was released for research use by the community at large.

EVALUATION RESULTS

The evaluation results are given in the table below in terms of the highest score approached by the best system to the nearest percentage point. Early on in some of the tasks, only recall (R) and precision (P) were calculated, but usually the combined F-measure (F) was used with recall and precision weighted equally.

The scores for Scenario Template did not ever get above the mid-60's for F. The reasons for this barrier are several and were partially addressed when multiple tasks were attempted at varying levels of processing. Scoring and template design issues that interfere with meaningful measures of progress are discussed in the next section and are being addressed in future evaluation methods.

The scores for Named Entity are in the mid-90's and are close to human performance. The machines and the annotators are still significantly different in their performance, but automation is preferable due to its speed. Applications of Named Entity have been successful in assisting humans in processing large amounts of textual data.

The scores for Template Element and Template Relations are also high enough to make the technology reliable for use by analysts. The Template Elements extracted from newswire articles are indicative of the content of the article for most purposes.

Although the coreference scores are lower than the Template Element scores, enough coreference is being processed to achieve reliable results in Template Element.

The multi-lingual scores are impressive both in Scenario Template and in Named Entity. Some of the variability is due to changes of domain between training and test documents in MUC-7. The results were surprising because many of the developers were not native or fluent speakers of the languages on which their systems were evaluated. Also, differences in style across

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

languages sometimes made processing easier in languages other than English.

EVALUATION ALGORITHMS

Evaluation Metrics

The evaluation metrics used for information extraction were adapted from Information Retrieval early in MUC-3 [1]. Both SGML markup in the text stream and template slot fills are scored automatically. The determination of scores for the coreference equivalence classes is based on a model theoretic algorithm that counts the minimal number of links that must be added to make the classes in the answer key and system response match [4].

Statistical Significance Testing

At the end of MUC-3, a statistical significance testing algorithm was developed to determine the significance of the results of the evaluation [3]. The method is a computer intensive method called approximate randomization and is based on a document-by-document comparison of performance for each pair of systems in the evaluation. The results are graphed and the sets of systems that are not significantly different from each other in performance on the test set are enclosed in the same circle. The method does not say that results are significant within a certain percentage, but rather looks at the characteristics of the performance of the systems across all documents.

Interannotator Scoring

During the course of the Tipster evaluations a method for measuring interannotator agreement was provided by the scoring software. This measure and the accompanying error reports assisted during the development of task definitions using training data and during the development of training and test datasets. The scoring software is designed to work domain-independently so it was easy to adapt for different scenarios and template slot designations in ST, TE, and TR and SGML markup for NE and CO. The key2key configuration option needed only be given to score an answer key against an answer key. Feedback is given in a strict fashion as to whether the annotators' keys agreed

on the alternatives and optional elements allowed only in answer keys.

User Interfaces

The scoring software has formatted reports and a GUI for viewing evaluation results in all languages. These tools for visualizing systems errors assisted developers in debugging their systems and in presenting their results. The user interfaces were designed based on input from the participants and the customers.

Remaining Scoring Issues

Alignment

The tree structure of the Scenario Templates requires choosing which objects to score against the objects given in the answer key. In order not to over penalize systems, alignment is done to optimize the F-measure. However, the optimization is not exhaustive and in some cases does not converge. Instead of scoring slot fills as both missing and spurious in less than ideal mappings, the scorer tends to map and score the mismatching slot fills as incorrect and the F-measure is calculated in such a way as to minimize the negative effect of missing and spurious slots.

In the future, the tree structure of the template will be greatly simplified so that the alignment problem is insignificant in understanding the results during development and testing.

Linchpin Effect

Another issue that arose in task design was the problem of penalizing a system in multiple places for one mistake. The inherent interdependencies of information especially in event descriptions made this aspect of task design difficult. Clearly over the course of Tipster, the annotations and templates changed to show the amelioration of this effect. Future evaluations will still need to beware of it.

CURRENT WORK

Currently, the Named Entity task used to evaluate information extraction systems on newswire articles is being adapted to evaluate such systems when processing errorful data, specifically the transcriptions done by speech recognition systems on Broadcast News. The purpose of this form of evaluation is two-fold. Improvements in speech recognition are expected to focus on the information-bearing elements of the signal. Improvements in information extraction are expected to make the systems more robust in the kinds of input they can handle.

FUTURE DIRECTIONS

Plans for the future of evaluations of information extraction from Broadcast News are now focused on event extraction. The template will be a simple set of slots for each event type and the fills will be unprocessed text extracted from the transcriptions. The fill rules will be simplified and the annotations will be more closely agreed upon. The alignment problem and the linchpin effect will be minimized.

ACKNOWLEDGMENTS

In a successful program such as Tipster there are many people to thank. The sponsors have been many over the years and none of these results would have been achieved without their efforts to keep the program on track. The participants in the evaluations have given their all in making information extraction emerge as an application-rich technology. The wider computational linguistics community responded with support and let their work be changed by evaluations. We are all grateful for this opportunity to have had an impact.

REFERENCES

[1] Chinchor, N.; Hirschman, L.; and Lewis, D. D. (1993). "Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3)." *Computational Linguistics*, 19(3),409 - 449.

[2] Merchant, R.; Okurowski, M.E.; and Chinchor, N. (1996). "The Multilingual Entity Task (MET) Overview." In *Proceedings, Tipster Text Program (Phase II)*. Morgan Kaufmann. San Mateo, CA.

[3] Chinchor, N. (1992). "The Statistical Significance of the MUC-4 Results." In *Proceedings, Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufmann. San Mateo, CA.

[4] Vilain, M.; Burger, J.; Aberdeen, J.; Connolly, D.; and Hirschman, L. (1995). "A Model-Theoretic Coreference Scoring Scheme." In *Proceedings, Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann. San Mateo, CA.