THE MULTILINGUAL ENTITY TASK (MET) OVERVIEW

Roberta Merchant
Mary Ellen Okurowski
Department of Defense
9800 Savage Road
Ft. Meade, MD 20755-6000

Nancy Chinchor
Science Applications International Corporation
10260 Campus Pt. Dr. M/S A2-F
San Diego, CA 92121

In November, 1996, the Message Understanding Conference-6 (MUC-6) evaluation of named entity identification demonstrated that systems are approaching human performance on English language texts [10]. Informal and anonymous, the MET provided a new opportunity to assess progress on the same task in Spanish, Japanese, and Chinese. Preliminary results indicate that MET systems in all three languages performed comparably to those of the MUC-6 evaluation in English.

Based upon the Named Entity Task Guidelines [11], the task was to locate and tag with SGML named entity expressions (people, organizations, and locations), time expressions (time and date), and numeric expressions (percentage and money) in Spanish texts from Agence France Presse, in Japanese texts from Kyodo newswire, or in Chinese texts from Xinhua newswire[1]. Across languages the keywords "press conference" retrieved a rich subcorpus of texts, covering a wide spectrum of topics. Frequency and types of expressions vary in the three language sets [2] [8] [9]. The original task guidelines were modified so that the core guidelines were language independent with language specific rules appended.

The schedule was quite abbreviated. In the fall, Government language teams retrieved training and test texts with multilingual software for the Fast Data Finder (FDF), refined the MUC-6 guidelines, and manually tagged 100 training texts using the SRA Named Entity Tool. In January, the training texts were released along with 200 sample unannotated training texts to the participating sites. A dry run was held in late March and early April and in late April the official test on 100 texts was performed anonymously. SAIC created language versions of the scoring program and provided technical support throughout.

Both commercial and academic groups participated. Two groups, New Mexico State University/ Computing Research Lab (NMSU/CRL) and Mitre Corp. elected to participate in all languages, SRA in Spanish and Japanese, BBN in Spanish (with FinCen) and Chinese, and SRI, NEC/University of Sheffield, and NTT Data in Japanese. Prior experience with the languages varied across groups, from new starts in January to those with considerable development history in multilingual text processing.

The MET results have been quite instructive from a number of different angles. First of all, multilingual named entity extraction is a technology that is clearly ready for application as the score ranges indicate in Table 1. Second, the informal anonymous nature appeared to encourage experimentation which is evidenced in the technical discussion of the summary site papers [1][6][12]. Third, system architectures have evolved toward increasing language portability [1][3][4][5][7], and, fourth, new acquisition techniques are accelerating development [1][4][5]. Fifth, resource sharing continues to play an important role in fostering technol-

Language	High Range	Low Range
Spanish	93.04	83.40
Japanese	92.12	70.79
Chinese	84.51	72.21

Table 1: MET Results

1. The language texts were supplied by the Linguistic Data Consortium (LDC) at the University of Pennsylvania.

ogy development. For example, two of the three sites in Chinese shared a word segmentor developed by NMSU/ CRL[1][4].

An additional contribution of MET was the baselining of human performance (Table 2). Dry run test data created by the language teams were analyzed to obtain consistency and accuracy scores as well as timing on the task. Analysts averaged eight minutes per article for annotation, including review and correction. Analysis revealed that inter-analyst variation on the task is quite low and that analysts performed this task accurately. This contrasts significantly with human performance data on a more complex information extraction task in MUC-5 [13]. When human baseline data are juxtaposed with the system scores, it is clear that the systems are approaching human accuracy with a much higher speed, offering further support for readiness for application.

Language	Consistency	Accuracy	
		High Range	Low Range
Spanish	92.92	91.42	88.62
Japanese	95	98	97
Chinese	94.32	98	95.94

Table 2: Inter-analyst Results

The scores in Tables 1 and 2 are the F-Measures obtained by the scoring software. The F-Measure is used to compute a single score in which recall and precision have equal weight in computation. Recall, a measure of completeness, is the number that the system got correct out of all of those that it could possibly have gotten correct; and precision, a measure of accuracy, is the number of those that it got correct out of the number that it provided answers for.

The F-Measures in Table 1 were produced by the automated scoring program. The program compares the human-generated answer key and the system-generated responses to produce a score report for each system. The low and high F-Measure scores from the formal test held in late April represent the current performance of the systems in this experimental evaluation.

The scoring software performs two processes: mapping and scoring. After parsing the incoming answer key and system response, it determines what piece of information in the response should be scored against each piece of information in the key. This process of alignment is called mapping and relies on the text being overlapping at least in part and, in cases where more than one mapping possibility exists, the software optimizes over the F-Measure for that piece of information. The scoring results are then tallied and reported.

The F-Measures in Table 2 for the human performance baseline were also produced by the automated scoring program. The consistency scores are the F-Measures resulting from comparing the two analysts' answer keys. The accuracy results are the F-Measures obtained by comparing each analyst's answer key against the final answer key. The measures are reported anonymously as a high and a low score.

In terms of the evaluation methodology, a number of lessons were learned from this experimental evaluation. The first was that the scoring software development effort would be improved by requesting realistic data from participants as early as possible for software testing instead of waiting until the dry run. An analysis of the order in which the data was provided, the timing of the distribution of the data, and the reliability of that data suggest that the results reported here are really the "floor" of what the technology is currently capable of rather than the "ceiling."Given that the systems are performing so close to human performance, it will be necessary to perform significance testing in the future. This testing will include human-generated responses in the test.

The Multilingual Entity Task section of this volume is a collection of papers that review the evaluation task and the participating systems. This overview paper is followed by three papers, discussing the task by language. Papers from each of the sites then briefly provide technical descriptions of their systems and participation in MET.

References

[1] Aberdeen, John, John Burger, David Day, Lynette Hirschman, David Palmer, Patricia Robinson, and Marc Vilain. "MITRE: Description of the ALEMBIC System as Used In MET" in this volume

[2] Anderson, Don A. "The Multilingual Entity Task A Descriptive Analysis of Enamex in Spanish" in this volume.

[3] Aone, Chinatsu. NameTag™ Japanese and Spanish Systems as Used for MET" in this volume.

[4] Ayuso, Damaris, Daniel Bikel, Tasha Hall, Erik Peterson, Ralph Weischedel, Patrick Jost. "Approaches in MET (Multilingual Entity Task) in this volume.

[5] Cowie, Jim. "CRL's Approach to MET" in this volume.

[6] Eriguchi, Yoshio and Tsuyoshi Kitani. "NTT Data: Description of the Erie System Used in MUC-6" in this volume.

[7] Kameyama, Megumi. "MET Name Recognition with Japanese FASTUS" in this volume.

[8] Keenan, Thomas. "An Interpretive Data Analysis of Chinese Named Entity Subtypes" in this volume.

[9]Maiorano, Steven and Terry Wilson, "Multilingual Entity Task: Japanese Results" in this volume.

[10] Sundheim, Beth. Proceedings of the Message Understanding Conference-6 (MUC-6), Morgan Kaufmann Publishers, Inc.: San Francisco, 1996.

[11] Sundheim, Beth. "Guidelines for Named Entity Task," Version 2.3, 1995.

[12] Takemoto, Yoshikazu, Takahiro Wakao, Hiroshi Yamada, Robert Gaizauskas, Yorick Wilks, "NEC Corporation and University of Sheffield: Description of NEC/Sheffield System Used for MET Japanese" in this volume.

[13] Will, Craig. "Comparing Human and Machine Performance for Natural Language Information Extraction: Results from the TIPSTER Text Evaluation - IDA" in Proceedings TIPSTER Text Program (Phase I), Morgan Kaufmann Publishers, Inc.: San Francisco, pp. 179 - 193, 1993.