# MULTILINGUAL ENTITY TASK (MET): JAPANESE RESULTS

*Steven Maiorano*
Office of Research & Development
Washington, D.C. 20505
E-mail: smaioran@ord.gov

*Terry Wilson*
Department of Defense
Ft. Meade, MD
E-mail: tsw@afterlife.ncsc.mil

## Introduction*

Japanese was one of the languages selected for evaluation of named entity identification algorithms in the TIPSTER-sponsored Multilingual Entity Task (MET) program. As with the Spanish and Chinese groups (Table 1), Japanese systems automatically marked the names of organizations, people, and places within entity name expressions (ENAMEX), dates and times within time expressions (TIMEX), and percents and money within number expressions (NUMEX). The participant Japanese systems were developed in a four-month period of time and output results comparable to the Message Understanding Conference-6 (MUC-6) [1] English language systems with F-Measures between 70 - 90% [2].

| Japanese | Spanish | Chinese |
|---|---|---|
| Mitre | Mitre | Mitre |
| NMSU | NMSU | NMSU |
| SRI | BBN/Treasury | BBN |
| SRA | SRA | |
| NEC/Sheffield | | |
| NTT Data | | |

**Table 1: MET Participants**

## The Corpus

Since MET was designed to tackle the MUC-6 named entity task in foreign languages, the government needed to acquire a corpus of articles rich in references to people, places, and organizations. A search of Kyodo newswire data using the keyword "記者会見" (press conference) yielded the desired 100-article development and 100-article test corpora. In the test corpus, 71% of the tags were of the ENAMEX type; that is, the tagged items were references to organizations, people, and places (Table 2). By contrast, for example, the 150-article TIPSTER Phase I test corpus contained only 75 instances of person names, or just 1% of all corpus tags [3].

| Tag Type | % of Test Corpus |
|---|---|
| ENAMEX | 71% |
| ORG | 25 |
| PERSON | 14 |
| LOC | 32 |
| TIMEX | 26 |
| DATE | 19 |
| TIME | 7 |
| NUMEX | 3 |
| PERCENT | 2 |
| MONEY | 1 |

**Table 2: Distribution of Tag Types**

## Human Performance

One motivation for conducting the named entity task in a foreign language such as Japanese was to promote techniques for tackling language-specific difficulties in recognizing the names of people and organizations. Unlike English, Japanese cannot rely upon orthographic clues like capitalization to identify proper nouns. For this reason, and based upon the authors' own manual tagging experience, we felt that identification of ENAMEX types would be the most challenging to the participant systems. (See Hard Tag Type below)

The government MET Japanese team was accorded the opportunity to test this hypothesis during the course of preparing dry-run keys for the initial systems test in early April. Each of us manually tagged the same 25 articles, then looked at the resulting annotator variation. The discrepancies between the two sets of tagged data were discussed and resolved. The final product formed ground truth or the keys against which the automatic participant systems were scored. Each of our original manually tagged versions was also scored against the keys. The figures in Table 3 are similar due to the small number of articles and lack of degradation in human performance over the short period of time it took for each of us to tag (average: < 5 minutes per article). Nonetheless, the ORGANIZATION and LOCATION subtypes were the most prone to error.

---

* *This material has been reviewed by the CIA. That review neither constitutes CIA authentication of information nor implies CIA endorsement of the author's views.*

| Tag Type | Average F Measure |
|---|---|
| MONEY | 100% |
| TIME | 100 |
| DATE | 99.75 |
| PERSON | 99.75 |
| LOCATION | 97 |
| PERCENT | 96.5 |
| ORGANIZATION | 96 |

**Table 3: Human Performance**

Table 4 shows the group average F-Measures for the participant systems against both the dry-run and test keys. The government's intuitive assumption concerning the relative difficulty of identifying ENAMEX types -- people, places, and organizations -- was borne out.

| Tag Type | Dry Run | Tag Type | Test |
|---|---|---|---|
| MONEY | 90% | PERCENT | 96% |
| PERCENT | 88 | MONEY | 95 |
| DATE | 86 | DATE | 94 |
| LOCATION | 81 | TIME | 93 |
| PERSON | 71 | LOCATION | 82 |
| TIME | 71 | PERSON | 77 |
| ORG | 62 | ORG | 73 |

**Table 4: Systems Performance**

## Easy Tag Types: NUMEX & TIMEX

While the entity name expressions were relatively difficult to handle, the number (NUMEX) and time (TIMEX) expressions encompassing the tag subtypes PERCENT & MONEY and TIME & DATE, respectively, were handled proficiently by the participant systems. As Table 4 shows, the group average F-Measure for these tag types was over 90% on the test data.

PERCENT

As in English, the typical Japanese contextual pattern for generating a valid PERCENT tag was an Arabic numeral + the "%" sign, e.g., "70%." The fact that the systems collectively scored less than 100% (96%) indicates that this pattern was not universal. Indeed, the Japanese development and test articles represented percentages in various other ways such as an Arabic numeral + the kanji (Chinese character) for percent " 7 割;" an Arabic numeral + the katakana (foreign loan-word script) for percent, e.g., " 7 0パーセント;" and a kanji numeral + the kanji for percent, e.g., "七割." In addition, the MET Japanese Guidelines [4] stipulated that fractions such as " 1 0分の一" (1/10), which are easily calculable as percentages, should also be identified. Although the above-mentioned patterns are more varied than what one typically encounters in English texts, they nevertheless constitute a standard finite list which the participant systems processed well.

MONEY and TIME

The contextual patterns manifested in the development and test corpora for representing MONEY and TIME were also limited. Typically, MONEY was specified by an Arabic numeral + a monetary unit in katakana, e.g., " 2 億ドル" ($200 million). The occasional mistake made by the systems involved not identifying a monetary unit other than the predominant dollar or yen such as "ポンド" (British pound).

TIME expressions were also straightforward in their manner of representation. Examples of valid tags included "0930," "朝" (morning), "午後 5 時" (5 PM), etc. An anomalous string such as "午前零時すぎ" (past midnight), in which the kanji "零" was used rather than the numeral "0," caused problems for some systems.

DATE

The Japanese participant systems processed DATE expressions successfully despite the demands made by the MET Japanese Guidelines [4] concerning what should be tagged and the wealth of patterns used to represent those expressions. In addition to absolute DATEs such as " 2 6 日" (26th) and "火曜日" (Tuesday) MET Guidelines stipulated that relative DATEs such as "来年 7 月" (next year July) were to be identified as well.

The requirement for tagging relative dates, furthermore, introduced to this task a class of Japanese semantic attachments that complicated the identification process. Whereas "来" (next, coming) and "去" (last) are in the initial position of the phrases "来年" (next year) and "去年" (last year), other semantic attachments such as "末" (end of) or "初め" (early, beginning of) , which add semantic content to the DATE expression and are, therefore, integral parts of the DATE tag, are in the postposition of phrases. Examples are "月末" (end of the month) and " 5 月初め" (early May). It was not uncommon, therefore, to encounter in these texts semantic-laden relative DATE expressions containing both prepositions and postpositions, e.g., "来年 3 月末" (end of March next year). In the end, the Japanese systems identified DATEs with an extensive number of different semantic attachments at the average rate of 94%.

## Hard Tag Type: ENAMEX

LOCATION

LOCATION expressions were typified by entities that likely would be contained in a gazetteer or similar on-line resource. References to "イスラエル" (Israel) or "米国" (U.S.) for example were readily identified by the systems. Other semantic clues such as the locative designators "県" (prefecture) or "州" (state) assisted in recognizing more obscure place names.

This task was complicated, however, by the prevalence of embedded LOCATION elements within ORGANIZATION expressions and the effects of context upon tag type. References to LOCATION frequently

appeared within phrases that might or might not subsume the LOCATION under another tag. For instance, "U.S.-Japan trade negotiations" would be an event not captured by a singular tag, but by two LOCATION tags for U.S. and Japan. However, the reference to "米" (U.S.) in "米国務省" (U.S. Dept. of State) was considered an integral part of the ORGANIZATION name and not, therefore, segmented and tagged separately. Determining when to segment and not segment the place sub-component was a complicating factor in producing the proper tag type. Furthermore, a correctly identified ORGANIZATION such as "国会" (Diet) was directed by the Guidelines to be tagged as LOCATION when the context in which it was used indicated that the Diet was a facility or structure -- i.e., if a press conference were being held there. The inability of systems to handle this complex contextual shift lowered the group F-Measure average for LOCATION.

PERSON

Although there is no ready-made on-line resource for person names like a gazetteer for place names, (and even if there were, its enormous size would slow substantially processing speed), valid PERSON expressions often contain people designators in the form of titles, roles, or positions such as "氏" (Mr.), "社長" (chairman), or "大統領" (president) like in "ムバラク大統領" (President Mubarak). And, although the number of different designators manifested throughout the corpora was sizable, the participant systems identified well person names occurring in this type of pattern.

There were, however, other patterns in evidence which conflicted with the predominant one. Preeminent among these alternate patterns was the expression in which a title/position designator was preceded by both a PERSON and LOCATION, e.g., "ムバラクエジプト大統領" (literally: Mubarak Egypt President). Systems typically tagged the entire phrase preceding the title -- Mubarak Egypt -- as PERSON, or Egypt alone as LOCATION. Therefore, the person name was either mistagged or not tagged at all.

ORGANIZATION

Identifying and categorizing complex noun phrases in strings where there is no capitalization and whitespace make this type of expression the most difficult to process (group F-Measure average 73%). Normally, the corporate designator "社" (Co., Corp.) would assist in identifying an ORGANIZATION. However, the MET domain focused on political rather than commercial entities, so there were very few instances of this designator. And, although bureaucratic descriptors like "省" indicate Japanese ministries, often well-known ministries such as "通産省" (MITI) are aliased (通産) without mention of the canonical form.

In addition, the most prevalent entities properly identified as ORGANIZATION in these texts included groups, offices, labs, etc. -- that is, noun phrases which could be proper nouns depending upon context. For example, "宮田工場" could be the name of a particular factory, Miyata Factory, or a generic factory located in Miyata; similarly, "新兵庫銀行" could be the New Hyogo Bank, the new (e.g., rebuilt) Hyogo Bank, or a new Hyogo Bank (i.e., one bank in the Hyogo Bank chain).

To complicate matters further, once a complex NP like "通産電気通信小委員会" (MITI Telecommunications Subcommittee) is determined to be a proper noun, the systems next were required to tag as ORGANIZATION each constituent part of the hierarchical relationship expressed within the phrase. In this case, there were two: MITI (parent) and Telecommunications Subcommittee (child).

## Summary

The Japanese systems showed excellent overall results despite a very compressed development cycle. They handled comparatively easy types of expressions with a high -- >90% -- degree of accuracy, and the hard expressions with surprising proficiency, thereby promising marked improvement in the near term and the capability to work in conjunction with other language processing technologies such as Machine Translation (MT) and text summarization [5].

## Acknowledgements

The authors wish to thank Adina Miller and Wiley Harris of ORD and Ron Dolan of the Library of Congress for their help in the organization and presentation of the data provided in this paper.

## References

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

[2] Merchant, Roberta, Okurowski, M. E., and Chinchor, N.; "*The Multilingual Entity Task (MET) Overview*;" in this volume.

[3] Proceedings TIPSTER Text Program (Phase I); Morgan Kaufmann Publishers, Inc.; San Francisco, CA; 1993.

[4] Maiorano, Steven; "*Multilingual Entity Task (MET) Definition for Japanese*;" March 8, 1996.

[5] Aone, Chinatsu, Maiorano, S., Okurowski, M.E.; "*Requirements and Applications of Thing Extraction*;" presentation at TIPSTER Text Program (Phase II) 12-Month Workshop; Chantilly, VA; May 19, 1995.