DOMAIN AND LANGUAGE EVALUATION RESULTS

Mary Ellen Okurowski
Department of Defense
9800 Savage Road, Fort Meade, Md.20755
meokuro@afterlife.ncsc.mil

INTRODUCTION

The Fifth Message Understanding Conference (MUC-5) focused on the task of data extraction for two distinctly different applications, one within the domain of joint ventures (JV) and the other within the domain of microelectronics (ME). For each application, the task could be performed in either English and/or Japanese, giving four combinations: English Joint Ventures, Japanese Joint Ventures, English Microelectronics, and Japanese Microelectronics.

Interpreting the evaluation results across domains and within a single domain between languages is affected by a number of factors. Differences in task focus, complexity, and domain technicality make it impossible to apply inferential statistics between domains. In addition, even though the task and the template design were the same across languages within a single domain, differences in the types of text sources for each language and accompanying variations in template fills and fill rules by language also make it impossible to apply inferential statistics between the language pairs. Moreover, there is considerable variation in the participants' level of effort and funding, and not all of the participants worked in multiple languages and/or multiple domains.

In light of these factors, I will present descriptive statistics comparing error per response fill to address the following questions: (1) For both languages, what is the performance difference between domains? (2) Between domains, what are performance differences for the single shared object and for unattempted slots? (3) For both domains, what is the performance difference between languages? (4) For a single domain, what are representative differences at object and slot levels between English and Japanese? The discussion of domain and language differences will center upon general factors that influence performance in information extraction: the information defined for extraction, the information available in a corpus for extraction, and the way in which information is presented within a text.

DOMAIN PERFORMANCE

Summary scores by domain for error per response fill (official All-objects scoring) averaged for MUC-5 sites in Table 1 indicate a slightly better performance in microelectronics than in joint ventures for both languages. [1]This performance characteristic is also reflected in the individual languages within domains in the summary of language performance in Table 4.

LANGUAGE	JOINT VENTURE		MICROELECTRONICS	
	AVERAGE	RANGE	AVERAGE	RANGE
ENGLISH/JAPANESE	74	54-84	70	58-86

TABLE 1: ERROR AVERAGE/RANGE BY DOMAINS FOR MUC-5 SITES

[1] Averaged scores reported in this paper for MUC-5 sites are based upon 12 English JV, 5 Japanese JV, 7 English ME, and 4 Japanese ME sites. Scores for the GE/CMU TEXTRACT system and the TRW DEFT system are not included.

This performance difference by domain can be interpreted by examining differences between the domains in **the information defined for extraction.** Domain differences for the only object type, *ENTITY*, that was defined for both domains will be presented first, followed by differences for unattempted slots, i.e. those slots that some or all systems left unfilled.

Shared Slots in *ENTITY* Object

Defining information extraction tasks entails identifying (1) the pieces of information to be extracted, (2) how the pieces are related, and (3) how those pieces are to be represented in a database. The two domains in MUC-5 define different tasks and so vary along those three parameters, which are collectively called the "reporting conditions." This variation in reporting conditions must be taken into account when examining results for a shared object extracted in two different domains. For example, in scoring performance for the *ENTITY* object in the JV and ME domains, what is being evaluated is not just how systems extract entities, but how systems extract entities given the reporting conditions of the domain. Whereas in the JV domain, systems mainly extract principals in tie-ups or newly formed companies, in the ME domain, they extract entities in terms of their relation to processes as developers, manufacturers, distributors, purchasers, or users of microelectronics technology.

Table 2 presents the error per response fill for the shared slots for the *ENTITY* object for the three TIPSTER sites that participated in both languages for both domains. These scores have been averaged across the two languages. In general, sites have a slightly better performance for the four slots in the JV domain than in the ME domain. The effect of reporting conditions for the two domains may be evident here. In the JV domain, the identification of the tie-up event is the central task.There are only two role distinctions to be made for the entities involved as either a principal in the tie-up or a newly formed joint venture company. In contrast, in the ME domain, the identification of the process and its attributes is the central task. Entity recognition, though pre-requisite to instantiate an ME capability, is actually in many ways auxiliary to the ME process itself. In addition, the entity must be assigned one of four different roles (developer, manufacturer, distributor, or purchaser/user), where no one slot dominates in terms of the number of expected fills. Thus, the entity recognition task in the ME domain is in some ways harder than in the JV domain.

	BBN		GE/CMU/MM		NMSU/BRANDEIS	
	JV	ME	JV	ME	JV	ME
NAME	52	55	41	52	58	58
LOCATION	93	94	61	72	80	79
NATIONALITY	83	89	68	78	79	95
TYPE	44	52	35	47	51	54

TABLE 2: ERROR ENTITY OBJECT SLOTS AVERAGED ACROSS LANGUAGES FOR THREE TIPSTER SITES.

Unattempted Slots

The effect of differences in the information defined for extraction on performance differences between the two domains can also be examined by reviewing unattempted slots. [2]Here, unattempted slots are defined as slots where actual is 0 and possible is greater than 0. Although a wide range of factors affect whether a site attempts a particular slot (e.g., its difficulty, fill frequency, clarity of definition in fill rules, or stability of definition in fill rules versions),

[2] This approach views a task independent of evaluation and its affect upon system development strategies. It ignores the fact that some objects appear more often and therefore contribute more to evaluation scores, which may shape the development efforts of sites seeking to maximize their scores by concentrating on high pay-off slots while ignoring slots with little pay-off in scoring.

for this discussion I will take the position that for each application a task is defined in terms of a certain number of objects with slots. A task requires a certain amount of work and each slot receives development effort. In the JV domain, there are ten objects with a total of 44 scored slots; in the ME domain there are nine objects with a total of 44 scored slots.

Review of unattempted slots in each domain allows us to determine how sites redefine the task in each domain by eliminating some subset of objects or slots from the task. Table 3 below indicates the task reduction averaged for MUC-5 sites for the two domains, calculated for each site by dividing the number of unattempted slots by the total number of slots. Even though performance differences for the ***ENTITY*** object indicate a somewhat better entity recognition for the JV domain (given reporting conditions) than for the ME domain, clearly, sites in the JV domain are more likely to reduce the task definition regardless of language. In both languages for JV, sites mainly redefine the task either by not filling slots in the Activity, Facility, Revenue, and Time Objects, or by not instantiating the object at all. In both languages for the ME domain, sites redefine the task mainly by not filling a subset of slots in the Etching, Packaging, and Equipment objects. There are no cases in ME where an entire object is not attempted.

This discrepancy in the extent of task redefinition between domains offers evidence of differences in task complexity that help us interpret the performance differences between domains. The greater likelihood for sites in the JV domain to eliminate slots and/or objects offers support to the view that the task is more complex for the JV domain than for the ME domain. The JV template design is a more complex structure, with a deeper set of embedded objects. Most of the unattempted JV slots are in the more deeply nested objects. The exception to this, the Activity object, is not part of the core template task for JV.

Discrepancies in development effort between domains for the TIPSTER sites further support the apparent greater complexity of the JV task. Notwithstanding the later start date for the ME domain and the more drastic revision process for the JV domain, all of the TIPSTER sites reduced the JV task more than the ME task. Moreover, the fact that three of the four sites working in both domains estimate that they expended considerably more development effort on JV than ME may further support that view.

DOMAIN	ENG/JPN	ENG		JPN	
	AVG	AVG	RANGE	AVG	RANGE
JOINT VENTURE	29	35	2-70	17	9-20
MICROELECTRONICS	13	17	0-30	6	0-12

TABLE 3: TASK REDUCTION (PERCENT) BY DOMAIN/LANGUAGE FOR MUC-5 SITES

LANGUAGE PERFORMANCE

Summary scores by language for error per response fill (official All-Objects scoring) averaged for MUC-5 sites in Table 4 indicate a better performance in Japanese than in English for both domains. This performance characteristic is also reflected in the individual domains.

LANGUAGE	JV/ME	JOINT VENTURE	MICROELECTRONICS
ENGLISH	75	77	73
JAPANESE	65	66	65

TABLE 4: ERROR SUMMARY AVERAGE BY LANGUAGE FOR MUC-5 SITES

This performance difference by language can be interpreted by analyzing how **information availability** (i.e., the amount of data of a given kind in the text that can be extracted) and **information presentation** (i.e., the manner in which different kinds of information are expressed) reflect the similarities and differences in evaluation results between the two languages. Language differences by object and slot will be presented first for the ME domain and then for the JV domain.

Microelectronics Domain: Impact of Information Availability

Evaluation results from the ME domain illustrate how the amount of information available in the corpus affects performance. In the ME domain, the application is directed at tracking microelectronics capabilities as evidenced in advances in four specific chip fabrication processes (lithography, layering, etching, and packaging). Identifying one of these processes associated with some entity triggers the tracking. Each of the four process objects is composed of a set of process-specific slots as well as a set of process-general slots shared by all four objects--Type, Device, Equipment slots. Error per response fill averaged by language for MUC-5 sites for the four process objects and their slots is presented in Tables 5-8.

LITHOGRAPHY SLOT	ENG	JPN
TYPE	66	58
DEVICE	78	83
EQUIPMENT	69	58
GRANULARITY	94	89

TABLE 5: ERROR FOR LITHOGRAPHY OBJECT BY LANGUAGE AVERAGED FOR MUC-5 SITES

LAYERING SLOT	ENG	JPN
TYPE	58	52
DEVICE	87	82
EQUIPMENT	76	69
FILM	80	97
TEMPERATURE	88	75

TABLE 6: ERRORFOR LAYERING OBJECT BY LANGUAGE AVERAGED FOR MUC-5 SITES

ETCHING SLOT	ENG	JPN
TYPE	71	75
DEVICE	84	74
EQUIP	75	57

TABLE 7: ERROR FOR ETCHING OBJECT BY LANGUAGE AVERAGED FOR MUC-5 SITES

ETCHING SLOT	ENG	JPN
ETCHANT	83	---
FILM	86	78
GRAN	91	86
TEMP	100	----

TABLE 7: ERROR FOR ETCHING OBJECT BY LANGUAGE AVERAGED FOR MUC-5 SITES

PACKAGING	ENG	JPN
TYPE	68	53
DEVICE	83	67
EQUIPMENT	88	---
PITCH	94	---
MATERIAL	74	73
PL/COUNT	62	34
UNITS	70	---
BONDING	87	---

TABLE 8: ERROR FOR PACKAGING OBJECT BY LANGUAGE AVERAGED FOR MUC-5 SITES

For both languages error per response fill is considerably lower for most of the process-general slots than for the process-specific slots. This discrepancy can be traced to the fact that similar types of information for the process objects are available in both languages and similar development strategies are employed: Emphasize high-frequency slots and de-emphasize low-frequency slots. Process-specific slots contribute significantly less to the total object scores than do process-general slots. In EME, process-specific slots in lithography, layering, and etching only comprise around 25% of the total object, and in JME for the same objects less than 20%. The same frequency pattern occurs in the development data; process-specific slots have a lower frequency of occurrence than the process-general slots. Note also that *DEVICE* and *EQUIPMENT* slots are pointers to other objects (with more slots) and the *TYPE* slot is a required slot that is indicative of the actual existence of a particular process within a text. Since information is more likely to be available for the process-general slots in both languages, more effort is directed at these higher pay-off slots than for the process-specific slots.This accounts for the better performance on process-specific slots in both languages.

But what accounts for the better Japanese than English performance in the EME domain? The Packaging object provides the first clue--differences in the amount of information between Japanese and English. Table 8 indicates that no test data are available for three of the process-specific slots for Japanese. In comparison to Japanese, the number of possible fills in the test set for English is considerably higher on all slots, not just these three; even factoring in the ratio of Japanese to English articles cannot account for this discrepancy. The development data also reflect this difference. Even though the task remained constant for these two languages in this domain, the type of data available for extraction for the Packaging object obviously differed for the two language pairs. There were simply fewer extractable items and thus fewer opportunities for error.

The amount of extractable items within a text affects the degree of difficulty of managing extractable items. In a single text, managing all the data elements associated with different multiple processes (i.e. multiple ME capabilities) is more difficult than managing only data elements associated with a single process. The English test set contained a higher proportion of texts with multiple processes (44%) than the Japanese test set (31%).The test sets also differed in the distribution of the types of multiple processes occurring with a text, e.g. whether a single text contains multiple processes of the same type or a combination of different process types. For the subset of texts containing multiple processes, Table 9 compares the percent distribution for the types of multiple processes. While the Japanese test set is more likely to contain a text with multiple layering or multiple lithography processes, the English test set is more likely to contain multiple packaging processes and combinations of process types.The average number of processes within a single text for layering, lithography, and etching types differs little across languages, but texts with multiple packaging processes contain twice as many processes in English as in Japanese.

LANGUAGE	PROCESS TYPE				MULTIPLE PROCESS TYPES
	MULTIPLE LAYERING	MULTIPLE LITHOGRAPHY	MUTLIPLE ETCHING	MULTIPLE PACKAGING	
ENGLISH	13	30	3	26	28
JAPANESE	23	42	3	13	19

TABLE 9: PERCENT OF MULTIPLE PROCESSES IN TEST SUBSET

Table 10 compares the average error per response fill for texts containing multiple processes with texts containing a single process for TIPSTER sites.With the exception of the GE/CMU/MM performance in Japanese, performance for all sites on texts with multiple processes was lower than on the texts with a single process.That a higher percentage of English test set texts contained multiple processes negatively affected the performance. Even though the Japanese test set contained more texts with multiple processes of the same type (and, in fact, an average lower performance on those text than on texts with multiple process types), the effect was ameliorated by the lower distribution of texts containing multiple processes in the Japanese test set. In other words, the greater likelihood of multiple processes within a single text in English (i.e. greater amount of extractable items) and the accompanying data management problems contributed to the weaker performance in English.

SITE	ENGLISH		JAPANESE	
	MULTIPLE PROCESSES	SINGLE PROCESS	MULTIPLE PROCESSES	SINGLE PROCESS
BBN	63	57	71	60
GE/CMU/MM	59	53	47	46
NMSU/Brandeis	74	67	63	54

TABLE 10: AVERAGE ERROR FOR MULTIPLE PROCESES VS. SINGLE PROCESS IN A SINGLE TEXT FOR TIPSTER SITES

Joint Ventures Domain: Impact of Information Presentation

Evaluation results from the JV domain illustrate the impact of information presentation. The way in which information is expressed in a single domain for two languages may differ. If texts in one language are more or less formulaic in structure and represent domain concepts in more or less standardized ways, then the texts in that language are more homogeneous in terms of discourse structure and terminology. As a result, texts in that language may be more easily exploited for information extraction than a more heterogeneous text corpus in a different language, even though the domain and application are the same. This appears to be the case for the JV domain for Japanese and English.

In the JV domain, the application is directed at tracking tie-ups among entities. Identifying entities engaged in some business activity in a tie-up relationship triggers the tracking. Error per response fill data by language averaged for MUC-5 sites for slots indicative of this identification are presented in Table 11

JOINT VENTURE OBJECT	SLOT	ENG	JPN
Template	Content	51	40
Tie-up Relationship	Entity	67	53
Entity	Name	71	49
Entity Relationship	Entity1	70	53
	Entity2	88	86
	Relationship2-1	76	46
Industry	Type	77	69

TABLE 11: ERROR JV SLOTS BY LANGUAGE AVERAGED FOR MUC-5 SITES

In general, the performance characteristic of lower Japanese error per response fill is consistent across the slots in Table 11. Systems perform better in Japanese in identifying the tie-up itself, participants in the tie-up, their relationship, and industry of the tie-up activity. This performance characteristic appears to be the result of the way in which information is presented in the Japanese text.

Preliminary analysis of the Japanese test set indicates that 60 percent or more of the articles have a prototypical text structure and that structure lends itself to a proficient extraction methodology.[3] Typically, an article contains one tie-up, and the relevant tie-up occurs in the first few sentences. Moreover, the tie-up signal is characterized by a stereotypical pattern as defined below:

X wa Y to....teikei shita to....happyo shita

In this pattern, X and Y are tie-up principals with the verb phrase "teikei shita" indicating a tie-up relationship.The key element is the topic marker "wa." That marker sets the stage for the entity to be the protagonist throughout the text and, in fact, for any other tie-ups mentioned in the article where only one of the entities is named. This prototypical structure gives the Japanese systems a headstart by providing a pattern into which missing or seemingly irrelevant information may later be inserted. In short, the presentation of the information in Japanese may facilitate extraction fills throughout the template and therefore may lead to better overall performance.

CONCLUSION

In general, evaluation results indicate a slightly better performance in the ME domain than in the JV domain and a better performance in Japanese than in English in both domains. This paper interprets these differences in terms of three formative factors in information extraction: information definition, information availability, and information presentation. Understanding performance differences from this perspective helps focus the examination on the information extraction problem rather than on dangerous, application-independent generalizations about domains and languages.

ACKNOWLEDGEMENTS

I would like to acknowledge invaluable database support from Terry Hand at the Department of Defense, in providing the data that formed the backbone of the arguments in this paper. I would also like to acknowledge insights

[3] Steve Maiorano, ORD, is preparing a paper to analyze system performance in the Japanese joint venture application to appear in the Proceedings from the TIPSTER Text Program, Phase One. His initial research has led to the definition of the prototypical structure and interpretation of its impact on evaluation results.

and critique from Beth Sundheim, NRaD, in editing this paper, and in particular in shaping my argument on the impact on performance of the amount of information.