CORPORA AND DATA PREPARATION

Lynn Carlson
Boyan Onyshkevych
Mary Ellen Okurowski
U. S. Department of Defense
Ft. Meade, MD 20755
email: {lmcarls, baonysh, meokuro}@afterlife.ncsc.mil

BACKGROUND

The data selection and data preparation efforts which led to the TIPSTER and Fifth Message Understanding Conference (MUC-5) evaluation corpora involved substantial effort, time and resources. The Government commitment to these selection and preparation efforts stems from four TIPSTER Program objectives: (1) to provide training data that would promote the development of information extraction technology, (2) to provide accurate test data to evaluate and baseline system performance in an objective manner, (3) to provide a baseline for human performance to understand and interpret machine performance, and (4) to support the larger Natural Language Processing community by making available a unique set of texts and templates in multiple domains and languages under ARPA support. This commitment was demonstrated through the managerial, technical, and administrative support to these efforts from various Government agencies, as well as through the contractual efforts with the Institute for Defense Analyses for data preparation and New Mexico State University for software tool development.

DOCUMENT CORPORA

Four language-domain pairs were used in the MUC-5 exercise, abbreviated as EJV, JJV, EME, JME to reflect the language (*E*nglish or *J*apanese) and the domain (*J*oint *V*entures or *M*icro*E*lectronics). Each of the four language-domain pairs has an associated set of 1200 to 1600 documents (a *corpus*), divided into the development set and the test sets. During the course of the TIPSTER program, up to three test sets were prepared for each language-domain pair, in addition to approximately 1000 development set documents for each corpus. These test sets, which were used for the TIPSTER 12-, 18-, and 24-month evaluations, ranged from 50 to 300 documents each. For MUC-5, the first test set was added to the development corpus, the second test set was used for the MUC-5 dry run, and the third test set was used for the MUC-5 evaluation. Selected from the overall pool in a random manner, the test sets reflect a similar distribution of sources, relevancy, and other document attributes as the development sets. There are a few exceptions, e.g., the first EJV test set does not contain documents from one of the sources added to the development and subsequent test sets.

These corpora consist of documents from a variety of newswire or newspaper sources, selected by a combination of automatic retrieval and manual filtering techniques. For example, the EJV corpus was retrieved from three text data sources (LEXUS/NEXUS, PROMT, and Wall Street Journal from ACL/DCI or TIPSTER Detection database CDROMs) by using traditional keyword-based document retrieval systems. These keywords for EJV included such stems as *joint venture, joint, venture, tie-up, collaborate, cooperate.* Though the majority of the documents were pulled by the keyword method, additional candidates were retrieved by random browsing through the corpora sources and identifying documents which appeared to be relevant. After a large pool of candidate documents was retrieved, these documents were manually scanned and separated into two groups: relevant or irrelevant. In order to test whether the Information Extraction systems were able to discriminate between relevant and irrelevant documents, the four corpora were then seeded with a certain number of irrelevant documents. The percentage of irrelevant documents functioning as "distractors" ranges from about 5% (for English Joint Ventures) to 30% (for Japanese Microelectronics). By comparison, the corpora used for previous MUCs used up to 50% irrelevant documents, stressing the document detection aspect of the task more strenuously than in TIPSTER/MUC-5.

The 200+ different sources used to build the English-language corpora include the *Wall Street Journal, Jiji Press, New York Times, Financial Times, Kyodo News Service*, and a variety of technical publications in fields such as communications, airline transportation, rubber & plastics, and food marketing. The Japanese-language sources used

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1993		**00-00-1993 to 00-00-1993**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Corpora and Data Preparation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Department of Defense,9800 Savage Road,Fort Meade,MD,20755**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**5**	

for the two Japanese corpora include *Asahi*, *Nikkei*, and *Yomiuri*.

Each document in the four development corpora has an associated filled-in template (see appendices to "Tasks, Domains and Languages" in this volume), representing the correct template or "answer key" that should be filled out for that document. The development corpora along with their associated templates were made available to the program participants during the course of the program.

TEMPLATE CORPORA

In order to provide the system developers with training data to illustrate the task and benchmark their development, filled-out templates for the approximately 1000 documents of each training set were provided as "keys". In addition, templates were produced for the initial TIPSTER program test cycles (12 and 18 months) and for the final joint MUC-5/TIPSTER (24 month) test. Table 1 provides the number of templates in each development and test set.

Table 1: Template Counts

	Development set	12-month test	18-month test / MUC dry run	24-month test / MUC evaluation	TOTAL
EJV	1000	100	200	282	1582
JJV	997	50	100	150	1297
EME	1000	0	100	300	1400
JME	850	0	50	100	1000
TOTAL	3847	150	450	832	5279

The templates were filled by experienced human analysts according to the same fill rules document (see below) and other supporting documentation that was provided to the system developers to define the exact syntax and semantics of the template fills.

FILL RULES

In addition to the template definition itself, which only defines the syntax in a BNF-like notation (see "Template Design for Information Extraction" in this volume), the analysts and participants in TIPSTER and MUC-5 were provided with the fill rule documents. At the highest level, the fill rules specify the reporting conditions for a given domain. These reporting conditions correspond to the general goals of the extraction task. For example, the fill rules document defines what information in a text constitutes evidence of a joint venture, and what minimal amount of supporting information is required in order to instantiate a template. Of note is that the conditions enumerated in the fill rules were determined from the document corpus and refined through the actual application of (earlier versions of) the fill rules to the corpus. At a more specific level, the fill rules delineate the conditions for instantiating an object, object by object, and for filling a slot, slot by slot. At the object and slot levels, the rules specify (1) what kind of evidence in the text is required for instantiation or fills and what, if anything, can be inferred, (2) the formatting conditions for data representation, and (3) the semantics of the data elements. Examples are often provided to highlight any one of these aspects.

The fill rules served as guidelines for two very different sets of users--the analysts and the system developers. Since the evolution of a fill rule document was driven to a large extent by its application to a text corpus, the analysts were key contributors to the fill rules in that they applied the rules and in so doing identified discrepancies, omissions, and exceptions to the rules. System developers, on the other hand, are mainly "consumers" of the rules, even though the TIPSTER participants did provide substantial input to the fill rules through questions and comments. Although reporting conditions as well as object and slot specifications need to be implemented in the extraction systems, the

developers of those systems also relied on the text corpus itself and analyst-filled templates to direct development.

In support of the fill rules document, other specialized documents were also provided, for example, expanding on the definition of a joint venture, or on the semantics of representing time expressions.

OTHER SUPPORTING MATERIAL

The Government also supplied on-line supporting materials to the analysts and the TIPSTER/MUC-5 participants. In many cases, this material was accessed to regularize or normalize the template fills. For example, the English language Gazetteer needed to be accessed in order to regularize geographic locational information. Compiled from a variety of sources, this resource provides place names for more than 240,000 locations around the world, along with type and containment information. For example, Baltimore is identified as a CITY, located in the PROVINCE (state) of Maryland, which is in the COUNTRY USA. The entire gazetteer entry for a location is used as the normalized fill for locational information in the template. Due to the small number of on-line geographic resources available for Japanese, a much more limited version of a Japanese gazetteer was manually produced by one of the Japanese analysts, with entries for all of the countries in the world, detailed listings for Japanese provinces, U.S. states, major cities for both countries, and other major cities worldwide that appeared in the JJV corpus. The Japanese language Gazetteer contains 1882 locations.

In the Joint Venture domain, the reporting of the products or business of the joint venture included classifying the product or service using the Standard Industrial Classification Manual compiled by the U. S. Office of Management and Budget. This resource contains a hierarchical classification of all the industry or business types in the U. S., for example, avocado farms, electric popcorn popper sales, management consulting. The template-filling task required that products or services be coded as a two-digit classification representing the second level in the hierarchy.

Other supporting resources for fill regularization include lists of currency names and abbreviations (e.g., the Dutch guilder is abbreviated NLG), lists of corporate abbreviations (like Inc, GMBH, or Ltd.) along with lists of countries where those abbreviations are typically used, and nationality adjectives (e.g., Iraqi, Irish). An additional set of resources was provided to system developers to assist in the extraction task, for example, lists of people's first names. All of these resources have been made available to the research community through the Consortium for Lexical Research at New Mexico State University.

DATA PREPARATION

The goal of data preparation was to have human analysts produce sets of development and test templates for each of the four corpora. The development templates served as models for system developers in the TIPSTER and MUC programs, and the test sets were used to measure system performance at six-month intervals (see above under TEMPLATE CORPORA). For each of the four domains, a group of experienced analysts was hired. These analysts met regularly over the course of 12 - 21 months (depending on the domain) to discuss domain and language-specific issues, iron out differences, and provide input to the fill rules, which evolved over time. The human analysts used a window-based tool for Sun Microsystems workstations, developed for the template-filling task by New Mexico State University's Computing Research Laboratory. One additional sub-task undertaken as part of the data preparation was the establishment of a performance baseline by measuring the performance of human analysts against each other and against the final "correct" version of various templates (see Table 2 below; for more detail, see also "Comparing Human and Machine Performance for Natural Language Information Extraction: Results for English Microelectronics from the MUC-5 Evaluation" in this volume).

Eleven of the nineteen analysts which comprised the four teams were hired by the Institute for Defense Analyses in Virginia (IDA); additional analysts from various Government facilities joined these teams. The Government technical management team (including the authors) led the effort to specify the domain, the definition of the templates, and the development of fill rules and other supporting materials, in addition to directing IDA, which was responsible for tracking template production and delivering prepared materials to the contractor sites, among other tasks.

In order to ensure maximal consistency and correctness in the analyst-produced keys, a variety of template-filling schemes were tried. Essentially, the schemes used different degrees of redundancy in producing each filled template, then used different methods to compare those template versions and to produce one final "most correct" version. Table 2 summarizes the different strategies that were tried. Most templates were produced using AB+B or

Table 2: Template Coding Schemes

Scheme Name	CODERS	CHECKERS	DESCRIPTION
A	A	-	One analyst produces template in one pass
A+A	A	A	One analyst codes, then checks it in a separate pass at a later time
A+B	A	B	One analyst codes template, another reviews it
AB+B	A and B	B	Two analysts independently produce codings, then one of them reviews both and produces composite version
AB+C	A and B	C	Two analysts independently produce codings, then a third analyst reviews those two and produces composite version.
ABCD +committee	each of A, B, C, D	all together	Each of four analysts produces coding independently, then final version produced by entire committee
ABCD + E	A, B, C, D + E	E	Each of four analysts produces coding independently, then final version produced by the fifth person

AB+C; JME was entirely produced using A+A. The analysts for the other three corpora rotated the A, B, C, and D positions among themselves. Even though redundant coding and checking methods were utilized, the templates that were produced could not be considered perfect; anomalies found by system developers were reviewed and changes were incorporated into the templates as appropriate.

TEMPLATE-FILLING STRATEGIES

The methodology used by the human analysts in filling templates was studied during the course of the task, partly to drive redesign of the tools and documentation to support the analysts' efforts. Although available resources did not permit extensive cognitive study of the mechanisms used by analysts, we did make some general observations about the strategies used by analysts.

A variety of approaches to template filling were used by the human analysts in filling out the templates. What follows is a characterization of the different strategies used by the five Japanese joint venture analysts (referred to as Analysts A, B, C, D and E) in analyzing the documents and filling out the corresponding templates.

The basic process can be divided into two parts: the start-up procedure and the actual template filling process, using the on-line tool. The start-up procedure includes both reading the text, and marking and annotating the hard copy. The template-filling process addresses the order in which the analysts actually filled out the objects and slots that represented the various pieces of information to be extracted from the text.

For the start-up procedure, three distinct approaches were identified. Scheme 1, used by two analysts, is charac-

terized by minimal marking of the hard-copy text before starting to code the template using the on-line tool. Analyst B would read the article twice through, then underline and label just the tie ups and entities before going to the tool. Analyst D would read and simultaneously underline entities and place check marks by other pertinent data; then he would begin coding.

In Scheme 2, also used by two analysts, a more detailed annotation of the hard-copy text was made. Analyst E would read through the hard-copy text and simultaneously underline and number entities, circle and number tie ups, and make other annotated comments, such as "E1 alias," "E2 official" (for alias or official associated with a particular entity). Moreover, this analyst would draw links between related pieces of information in the text, and would outline in the margins more complex objects, such as ACTIVITY, OWNERSHIP, and REVENUE. After this process was complete, the coding would begin. Analyst C's approach was similarly detailed, the only difference being that she would label all pertinent information using color-coded highlighters, e.g., green for ENTITYs, yellow for product/service strings, blue for FACILITY and TIME objects.

The third scheme, used by Analyst A, involved a mixture of initial marking, skimming, initial coding, annotating in detail, and then final coding. This analyst would read the beginning of the article, marking potential entities until a "tie-up verb" was found. Now certain that the article had a valid tie up, she would proceed to skim the remainder of the text, underlining or circling additional pertinent information. At this point, she would use the tool to code the initial portion of the template, i.e., the TIE-UPs, ENTITYs, and ENTITY-RELATIONSHIPs. After this key structure was in place, she would read through the remainder of the text, annotating in detail all potential product/service strings, and information about FACILITYs, REVENUE, OWNERSHIP, etc. Finally, the remainder of the template was coded using the tool.

Moving on to the template-filling process, a variety of breadth vs. depth-first strategies were used by the analysts. Four of the analysts would completely fill in all information about the first tie up before coding any additional tie ups. Analysts A, B, and E would fill in the TEMPLATE, TIE-UP, ENTITY, and ENTITY-RELATIONSHIP objects first. Then TIME, REVENUE, OWNERSHIP, PERSON and FACILITY objects were instantiated in no particular order. The ACTIVITY and INDUSTRY objects were filled in concurrently, usually last. This procedure was then repeated for additional tie ups. Analyst D followed a complete depth-first strategy for coding each tie up, filling in each slot in turn, so that if a slot pointed to another object, that object would then be filled in completely before proceeding to the next slot in the top level object. A breadth-first strategy for coding was used by Analyst C, who would fill in all tie-up objects and their respective entities first, and then code the remaining information for each tie up.

These varying strategies for annotating texts and coding templates did not seem to have a significant effect on the quality of the templates produced, and seemed to be a matter of personal preference. However, they give insight into the different ways in which humans approach a particular analytic task, and suggest that on-line analytic tools need to be sufficiently flexible to accommodate the styles of different human users.