PRC Inc.: DESCRIPTION OF THE PAKTUS SYSTEM USED FOR MUC-5

Bruce Loatman
Chih-King Yang

PRC Inc.
Technology Division
1500 PRC Drive
McLean, VA 22102
loatman_bruce@po.gis.prc.com

BACKGROUND

The PRC Adaptive Knowledge-based Text Understanding System (PAKTUS) was developed as an Independent Research and Development project at PRC from 1984 through 1992. It includes a core English lexicon and grammar, a concept network, processes for applying these to lexical, syntactic, semantic, and discourse analysis, and tools that support the adaptation of the generic core to new domains, primarily by acquiring sublanguage and domain-specific lexicon and topic patterns. The lexical, syntactic, and semantic analysis components required little adaptation for MUC-5, the most significant change being conversion of the task-specific semantic representations to object-oriented form. The discourse analysis component was modified to operate on the task-specific semantic structures, rather than the generic case frames.

APPROACH

The overall structure and operation of PAKTUS are shown in Figure 1. This is similar to the "generic system" described in [1]. Processing proceeds mostly sequentially, with the exception of the interaction between the syntactic and semantic components at the clause and noun phrase level, and between the lexical analysis and preprocessor.

For the MUC5 task, we added some bracketing capabilities to handle the special syntactic phenomena in the financial application domain which might cause problems for the parser or later extraction processes. These phenomena include company name, currency, temporal expressions, and one percentage expression. For example, "BRIDGESTONE SPORTS CO.", "BRIDGESTONE SPORTS TAIWAN CO" , "UNION PRECISION CASTING CO" and "TAGA CO." are recognized as company names during the bracketing phase; " 20 MILLION NEW TAIWAN DOLLARS" is bracketed as ((20000000 DOLLAR^CURRENCY BASE C^TAIWAN)); and "75 PCT BY" is treated as a preposition. The complete sentence parse rate for the MUC5 corpus was significantly improved by the bracketing process, minimizing the need for complex additions to the grammar.

Unlike the generic system, PAKTUS has no text filter or preparser. Full parses are attempted on all sentences, and the first syntactico-semantically successful parse of a sentence is accepted. Parse time is restricted as a linear function of the number of words, however, and parse fragments are returned, implicitly conjoined, if a full parse cannot be produced in the allotted time. Full parses were achieved for approximately 50 percent of the sentences in the MUC-5 corpus. Useful information was obtained from the fragmentary parses of the other half, however. Based on comparison of the MUC-5 error measures when fragmentary parses were included and excluded,

the fragmented parses yielded, on average, about one-third as much correct information as the full parses.

Another variation on the generic system is that fragment combination and semantic interpretation are integrated in PAKTUS, and semantic interpretation is divided into two distinct modules: one that produces a generic representation of the complete sentence, and a subsequent module that maps this into (possibly several) task-specific representations. In addition, at the lexical analysis phase, as mentioned above, each word was associated with both syntactic and semantic information. Lexical patterns were developed as an alternative to semantic interpretation based on full syntactico-semantic parses. This takes advantage of the information rich lexical analysis. A pattern matcher uses the results of both the lexical analysis (as in Figure 2a) and the syntactic analysis (as in Figure 3b) to extract information. This is invoked when the extraction based on full parsing fails to yield any data. The patterns are defined as regular expressions that are matched against the results of lexical analysis. When a match is found, the corresponding noun phrases (produced by the full parser) are extracted and the task-specific semantic representation is constructed from these.

PAKTUS has no separate lexical disambiguation module: that function is distributed across all system modules. Initially, words are assigned all possible meanings available in the lexicon. Senses that are inconsistent with any processing choice are filtered out when that choice is considered.

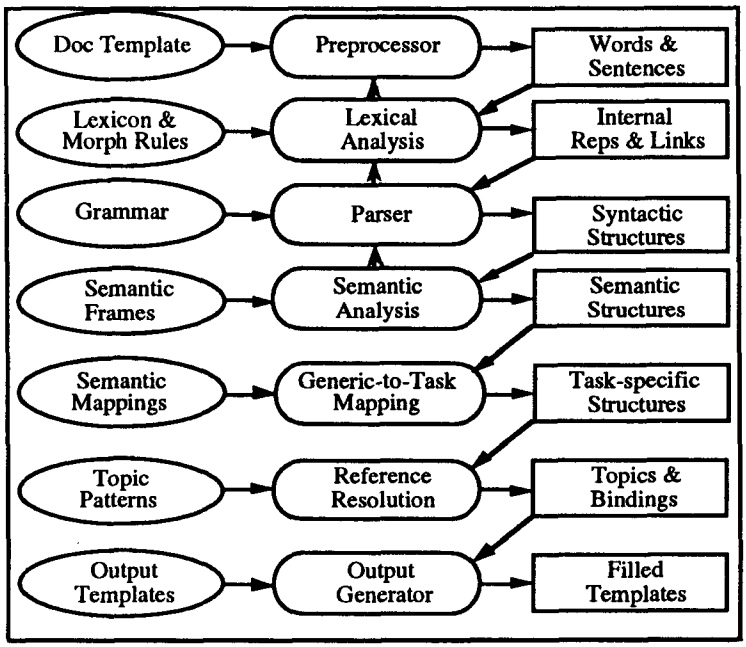

Figure 1: PAKTUS Modules and Control Flow

PROCESSING OF MUC-5 DOCUMENT 0592

Figure 2a shows the raw, unprocessed text of the first sentence (S1) of article number 0592, followed by its lexical analysis. This is the result of applying both the preprocessor and lexical analysis modules. Each word has one or more senses, represented as a root symbol, which is generally the concatenation of the English token, the "^" character, and the PAKTUS lexical category (e.g., "Set^Monotrans"), or as a simple structure involving a root, lexical category, inflectional mark, and sometimes a conceptual derivation (e.g., the structure "(Say2^Monotrans L^Monotrans S^ed)" represents the -ed form of one sense of "say"). For each word, all senses in

the PAKTUS lexicon are fetched or derived at this time; disambiguation is generally delayed until later phases. Many of the words are unknown to the PAKTUS lexicon; it will make guesses from the context. An example of an ambiguous word is "concern." Figure 2b shows some of the lexicon information for this word in PAKTUS. This includes a list of the 4 PAKTUS primitive words corresponding to the token "concern" plus objects containing information for each primitive word. These objects are embedded in a semantic network; they inherit much additional information from it.

Sample PAKTUS grammar specifications relevant to S1 are shown in Figure 3a, and the syntactic analysis of this sentence is shown in Figure 3b. This analysis is represented as a configuration of syntactic registers (the main ones are shown in the figure) and register fillers. The grammar fragment of Figure 3a recognized the bound clause in S1 ("... it has set up a joint venture...").

Several semantic frames that apply to S1 are shown in Figure 4a, and the generic semantic analysis of this sentence is shown in Figure 4b. PAKTUS represents the semantic analysis as five case frames: one for the sentence, one for each of the three subordinate clauses (only two of which are displayed in the figure), and one for the "joint venture" NP. The semantic rules are distributed in a network of objects like those in Figure 4a.

Information is organized in several hundred conceptual objects (e.g., C^CREATE - the concept for "set up" in S1) and case roles (e.g., R^RESULT - the thing created). Information about how to map from syntactic registers to case roles may be stored in concept frames, lexical frames, or role frames, along with semantic constraints on allowable fillers. For example, the R^RESULT role of C^CREATE is normally filled by the direct object (DO) register. This information is inherited by R^RESULT from the more general R^OBJECT role.

```
BRIDGESTONE SPORTS CO. SAID FRIDAY IT HAS SET UP A JOINT VENTURE IN TAIWAN
WITH A LOCAL CONCERN AND A JAPANESE TRADING HOUSE TO PRODUCE GOLF CLUBS
TO BE SHIPPED TO JAPAN.
        *** lexical analysis:
(((UNKNOWN-WORD L^COMPANY BASE C^UNKNOWN . "BRIDGESTONE SPORTS CO"))
  ((SAY^INTRANS L^INTRANS S^ED)  (SAY^TO-IO L^TO-IO S^ED)
   (SAY2^MONOTRANS L^MONOTRANS S^ED))
  (("17-NOV-89" L^TIME-DATE BASE)) (IT^NEUTER)
  ((HAVE^MONOTRANS L^MONOTRANS S^S)  (HAVE2^INTRANS L^INTRANS S^S)
   (HAVE^INTRANS L^INTRANS S^S)  (HAVE^HAVE L^HAVE S^S)
   (HAVE1^MONOTRANS L^MONOTRANS S^S))
  (SET^COLLECTION  (SET^MONOTRANS L^MONOTRANS S^ED)  SET^MONOTRANS)
  (UP^PARTICLE UP^PREP UP^DIRECTION) (A^DET)
  (JOINT\ VENTURE^ACTIVITY)
  (IN^PARTICLE IN^PREP) (TAIWAN^NATION)
  (WITH^PARTICLE WITH^PREP) (A^DET)  (LOCAL^SPACE-REL)
  (CONCERN^COPULA CONCERN^MONOTRANS  CONCERN^EMOTION CONCERN^BUSINESS)
  (AND^CONJ) (A^DET)
  ((JAPAN^NATION L^LANGUAGE BASE C^CHAR-OF)  (JAPAN^NATION L^ADJ BASE
    C^IT-BE-FROM)  (JAPAN^NATION L^INHABITANT BASE C^BE-FROM))
  ((UNKNOWN-WORD L^COMPANY BASE C^UNKNOWN  . "TRADING HOUSE"))
  (TO^PREP TO^PARTICLE) (PRODUCE^MONOTRANS)
  ((UNKNOWN-WORD VP BASE C^PRIMITIVE . "GOLF")
   (UNKNOWN-WORD L^COMMON BASE C^UNKNOWN  . "GOLF"))
  ((CLUB^GROUP L^GROUP S^S)
   (UNKNOWN-WORD L^INTRANS S^S C^UNKNOWN  . "CLUBS")
   (UNKNOWN-WORD L^MONOTRANS S^S C^UNKNOWN  . "CLUBS"))
  (TO^PREP TO^PARTICLE)
  (BE^BE BE^INTRANS BE^COPULA)
  ((SHIP^BITRANS L^BITRANS S^ED)  (SHIP^MONOTRANS L^MONOTRANS S^ED))
  (TO^PREP TO^PARTICLE) (JAPAN^NATION))
```

Figure 2a: Lexical Analysis of the First Sentence of Document Number 0592

```
"CONCERN"
(CONCERN^COPULA CONCERN^MONOTRANS CONCERN^EMOTION CONCERN^BUSINESS)
(CONCERN^COPULA (AKO (L^COPULA))
          (COMPLEMENT (WH-CLAUSE WH-TO-INF NP))
          (CONCEPT (C^BE-ABOUT))
          (ROLES (R^AFFECTED R^PROPOSITION R^FOCUS))
          (R^AFFECTED NIL (_ SUBJECT) (@ "INFO"))
          (R^FOCUS NIL (_ COMP) (@ (TRUE))))
(CONCERN^MONOTRANS (AKO (L^MONOTRANS)) (CONCEPT (C^MOTIV)))
(CONCERN^EMOTION (AKO (L^EMOTION)))
(CONCERN^BUSINESS (AKO (L^BUSINESS)) (TYPE (COUNT LEFT-ADJ-OF-N)))
```

Figure 2b: Some PAKTUS Lexicon Data Used for S1

```
(C-S-T1E (TO-STATE (E^)) (AKO (ARC))
         (INIT
          ((^.MAIN-VERB.LEX.HAS-FEATURE '(L^VERB ZERO-THAT))
           (*.MOOD_BOUND)))
         (RULE (L^S31RULE)) (LABEL (T1)) (FROM-STATE (C^))
         (NAME ("cS\\t1e")))
(L^S31RULE (AKO (L^SRULE)) (PRIORITY (0))
           (THEN NIL (ACTIONS (^.PROP_*))))
```

Figure 3a: Some PAKTUS Grammar Specifications Used for S1

```
(S (MAIN-VERB55 (SAY^TO-IO L^TO-IO S^ED))
 (SUBJECT53
  (NP (HEAD54
       (UNKNOWN-WORD L^COMPANY BASE C^UNKNOWN . "BRIDGESTONE SPORTS CO"))))
 (PROP11
  (T1 (MAIN-VERB51 (ESTABLISH^MONOTRANS L^MONOTRANS S^ED))
   (SUBJECT35 (NP (HEAD36 IT^NEUTER)))
   (DO14
    (NP (HEAD49 JOINT\ VENTURE^ACTIVITY) (DET50 A^DET))
    (PROP30
     (Z^ (MAIN-VERB47 PRODUCE^MONOTRANS)
      (SUBJECT35 (NP (HEAD36 IT^NEUTER)))
      (DO31
       (NP (HEAD44 (CLUB^GROUP L^GROUP S^S))
        (PROP32
         (Z^ (MAIN-VERB39 (SHIP^MONOTRANS L^MONOTRANS S^ED))
          (SUBJECT37 (NP (HEAD38 SOMEONE^SOME)))
          (DO35 (NP (HEAD36 IT^NEUTER)))
          (MODS40 (PP (PREP41 TO^PREP)
                      (PREP-OBJ33 (NP (HEAD34 JAPAN^NATION)))))))
         (DESC46  (UNKNOWN-WORD L^COMMON BASE C^UNKNOWN . "GOLF"))))))
    (MODS15
     (PP (PREP29 WITH^PREP)
      (PREP-OBJ18
       (NP (HEAD25 CONCERN^BUSINESS) (DET28 A^DET)
        (DESC27 LOCAL^SPACE-REL)
        (CONJ19
         (NP
          (HEAD20 (UNKNOWN-WORD L^COMPANY BASE C^UNKNOWN . "TRADING HOUSE"))
          (DET23 A^DET)
          (DESC22 (JAPAN^NATION L^ADJ BASE C^IT-BE-FROM)))))))))
   (MODS16
    (PP  (PREP17 IN^PREP)
     (PREP-OBJ12 (NP (HEAD13 TAIWAN^NATION)))))))
 (ADV56 ("17-NOV-89" L^TIME-DATE BASE)))
```

Figure 3b: Syntactic Analysis of S1

```
(C^CREATE (ROLES
           (R^AGENT R^RECIPIENT R^INSTR R^RESULT R^PURPOSE R^MATERIAL))
          (AKO (C^BEGIN))
          (R^MATERIAL NIL (_ (PREP-OBJ 'FROM^PREP))))
(R^RESULT (AKO (R^EFFECT)))
(R^EFFECT (KINDSOF (R^RESULT R^EVENT)) (AKO (R^OBJECT)))
(R^OBJECT
 (KINDSOF
  (R^AFFECTED R^EXPERIENCER R^COMPANION R^EFFECT R^PROPOSITION
              R^FOCUS R^PURPOSE R^MATERIAL R^RESISTANCE))
 (_ NIL (DEFAULT DO)) (AKO (PROP-ROLE)))
```

Figure 4a: Some PAKTUS Generic Semantic Specifications Used for S1

```
(C^ASSERT
 (R^AGENT53
  (F53
   (HEAD54 (UNKNOWN-WORD L^COMPANY BASE C^UNKNOWN . "BRIDGESTONE SPORTS CO"))))
 (R^PROPOSITION11
  (C^CREATE (R^INSTR35 (F35 (HEAD36 IT^NEUTER)))
   (R^RESULT14
    (C^ATTEMPT (HEAD49 JOINT\ VENTURE^ACTIVITY)
     (R^INSTR14 @F14)
     (R^COMPANION18
      (C^ACT (HEAD25 CONCERN^BUSINESS)
       (CONJ19
        (F19
         (HEAD20 (UNKNOWN-WORD L^COMPANY BASE C^UNKNOWN . "TRADING HOUSE"))))
       (CONJOINER24 AND^CONJ) (R^INSTR18 @F18)))
     (R^PURPOSE30
      (C^CREATE (R^INSTR35 (F35 (HEAD36 IT^NEUTER)))
       (R^RESULT31 (F31 (HEAD44 (CLUB^GROUP L^GROUP S^S))))))
   (R^PLACE12 (F12 (HEAD13 TAIWAN^NATION)))
```

Figure 4b: Generic Semantic Analysis of S1

Figure 5a gives an example of a semantic mapping rule. This consists of a pattern component, which in this case matches the semantic frame of Figure 4b, and a mapping specification (in the "slots" component). Figure 5b shows the task-specific semantic representation that results from applying this mapping to the generic semantic frame. The rule in Figure 5a is a refinement of one used in the final test MUC-5 system. The final test version recognized the tie up relationship in this sentence, but extracted only two of the three tie up entities. The new mapping rule better illustrates system features without significantly changing the output.

The pattern component specifies a semantic structure with a C^TALK concept as its root (the C^ASSERT concept shown in the semantic analysis of S1 is a specialization of C^TALK in PAKTUS), and with R^AGENT and R^PROPOSITION roles. The filler of the R^AGENT role will be bound to the pattern variable R^AGENT250. The R^PROPOSITION role must be filled by an instance of a C^BEGIN frame (the C^CREATE concept in S1 is a specialization of C^BEGIN), with a R^RESULT that is an instance of C^ATTEMPT (which joint venture is). If R^COMPANION or R^PURPOSE roles are filled in the C^ATTEMPT frame, information is extracted from them as well (they are optional, which is not indicated in the figure, but is marked in the actual mapping object).

The mapping portion of the rule specifies how to map the pattern variable bindings to a task-specific semantic object. This rule says that the result is a tie up relationship with tie up entities derived from the bindings of R^COMPANION248 and R^AGENT250, the joint venture taken as the filler of the R^RESULT role of the C^BEGIN frame, etc. The type of relationship (existing, former, etc.) is determined by the act-stage function, which computes the type from tense, aspect, and modality registers.

```
(P-TIE_UP_RELATIONSHIP143
 (AKO C^TALK)
 (R^AGENT ((> R^AGENT250)))
 (R^PROPOSITION
  (P-TIE_UP_RELATIONSHIP144
   (AKO ((CON250 IS C^BEGIN)))
   (R^INSTR ((< R^INSTR248)))
   (R^RESULT
    (P-TIE_UP_RELATIONSHIP150
     (AKO ((CON249 IS C^ATTEMPT)))
     (R^INSTR @P-TIE_UP_RELATIONSHIP150)
     (R^COMPANION ((R^COMPANION248 IS L^AGENT)))
     (R^PURPOSE
      (P-TIE_UP_RELATIONSHIP186
       (AKO ((CON248 IS C^CREATE)))
       (R^INSTR ((R^INSTR248 IS L^3RD-PERS-PRO)))
       (R^RESULT ((> R^RESULT248)))))))
   (R^PLACE ((R^PLACE248 IS L^LOCATION)))))
 (SLOTS
  (TIE_UP_RELATIONSHIP (TYPE (ACT-STAGE))
                       (TU-ENTITY R^COMPANION248 R^AGENT250)
                       (JOINT-VENTURE P-TIE_UP_RELATIONSHIP150)
                       (TU-ACTIVITY (INDUSTRY (I-TYPE PRODUCTION)
                                              (PRODUCT/SERVICE R^RESULT248))
                                    (SITE R^PLACE248)))
  (ENTITY (ENTITY-RELATIONSHIP (ENTITY1 R^COMPANION248 R^AGENT250)
                               (ENTITY2 P-TIE_UP_RELATIONSHIP150)
                               (RELATIONSHIP CHILD)))))
```

Figure 5a: Generic-to-Task-Based Semantic Mapping Rule

Figure 5b shows the task-specific representation that the mapping rule in Figure 5a produced when applied to the generic semantic representation of Figure 4b.

Another task-specific representation is shown in Figure 6, for the fourth sentence: THE NEW COMPANY, BASED IN KAOHSIUNG, SOUTHERN TAIWAN, IS OWNED 75 PCT BY BRIDGESTONE SPORTS, 15 PCT BY UNION PRECISION CASTING CO. OF TAIWAN AND THE REMAINDER BY TAGA CO., A COMPANY ACTIVE IN TRADING WITH TAIWAN, THE OFFICIALS SAID. This is shown to illustrate one of the limitations of the MUC-5 version of the system.

```
(SPEC35
 (ENTITY-RELATIONSHIP
  (SPEC37 (RELATIONSHIP "CHILD")
          (ENTITY2 "JOINT VENTURE")
          (ENTITY1  "LOCAL CONCERN")
          (ENTITY1 "JAPANESE TRADING HOUSE")
          (ENTITY1 "BRIDGESTONE SPORTS CO")))
 (TU-ACTIVITY
  (SPEC38 (SITE "TAIWAN")
          (INDUSTRY
           (SPEC39 (PRODUCT/SERVICE "GOLF CLUBS")
                   (I-TYPE "PRODUCTION")))))
 (JOINT-VENTURE ("JOINT VENTURE"
                 (ENTITY-RELATIONSHIP SPEC37)))
 (TU-ENTITY ("LOCAL CONCERN"
             (ENTITY-RELATIONSHIP SPEC37)))
 (TU-ENTITY ("JAPANESE TRADING HOUSE"
             (ENTITY-RELATIONSHIP SPEC37)))
 (TU-ENTITY ("BRIDGESTONE SPORTS CO"
             (ENTITY-RELATIONSHIP SPEC37)))
 (TYPE "EXISTING"))
```

Figure 5b: Task-Specific Semantic Representation of S1

```
(SPEC40
 (ENTITY-RELATIONSHIP
  (SPEC42 (ER-STATUS "CURRENT") (RELATIONSHIP "CHILD")
          (ENTITY2 "NEW COMPANY")
          (ENTITY1 "TAGA CO")
          (ENTITY1 "UNION PRECISION CASTING CO")
          (ENTITY1 "REMAINDER")
          (ENTITY1 "BRIDGESTONE SPORTS")))
 (TU-ACTIVITY
  (SPEC46 (SITE ("KAOHSIUNG" (APP "SOUTHERN TAIWAN")))))
 (OWNERSHIP
  (SPEC43 (OWNERSHIP-% "BRIDGESTONE SPORTS") (OWNERSHIP-% 75)
          (OWNED "NEW COMPANY")))
 (OWNERSHIP
  (SPEC44 (OWNERSHIP-% "UNION PRECISION CASTING CO") (OWNERSHIP-% 15)
          (OWNED "NEW COMPANY")))
 (OWNERSHIP
  (SPEC45 (OWNERSHIP-% "TAGA CO")
          (OWNED "NEW COMPANY")))
 (JOINT-VENTURE ("NEW COMPANY"
                 (ENTITY-RELATIONSHIP SPEC42)))
 (TU-ENTITY ("TAGA CO" (ENTITY-RELATIONSHIP SPEC42)))
 (TU-ENTITY ("UNION PRECISION CASTING CO" (LOC "TAIWAN")
             (ENTITY-RELATIONSHIP SPEC42)))
 (TU-ENTITY ("REMAINDER" (ENTITY-RELATIONSHIP SPEC42)))
 (TU-ENTITY ("BRIDGESTONE SPORTS" (ENTITY-RELATIONSHIP SPEC42)))
 (TYPE "EXISTING"))
```

Figure 6: Task-Specific Semantic Representation of S4

The complete filled templates for this article are shown in Figure 7. Some of the information, such as ownership percentages, that was extracted as shown in Figure 6, does not appear in the output templates. This is typical of the MUC-5 version of PAKTUS; we were unable to devote resources sufficient to complete the output generator component (the final processing module in Figure 1), so some information that was extracted was simply ignored by the final process.

```
<TEMPLATE-0592-84> :=
    DOC NR: 0592
    DOC DATE: 241189
    DOCUMENT SOURCE: "Jiji Press Ltd."
    CONTENT: <TIE_UP_RELATIONSHIP-0592-84>
    DATE TEMPLATE COMPLETED: 930820
<TIE_UP_RELATIONSHIP-0592-84> :=
    TIE-UP STATUS: EXISTING
    ENTITY: <ENTITY-0592-84>
            <ENTITY-0592-85>
            <ENTITY-0592-86>
            <ENTITY-0592-87>
            <ENTITY-0592-88>
    JOINT VENTURE CO: <ENTITY-0592-89>
    OWNERSHIP: <OWNERSHIP-0592-89>
    ACTIVITY: <ACTIVITY-0592-89>
<ENTITY-0592-84> :=
    NAME: CO
    TYPE: COMPANY
<ENTITY-0592-85> :=
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-89>
    NAME: UNION PRECISION CASTING CO
    LOCATION: TAIWAN (COUNTRY)
    TYPE: COMPANY
```

Figure 7: PAKTUS Template Fills for the Sample Document

```
<ENTITY_RELATIONSHIP-0592-89> :=
    ENTITY1: <ENTITY-0592-86>
    ENTITY2: <ENTITY-0592-89>
    REL OF ENTITY2 TO ENTITY1: CHILD
<ENTITY-0592-86> :=
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-89>
    NAME: TAGA CO
    TYPE: COMPANY
<ENTITY-0592-89> :=
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-90>
            <ENTITY_RELATIONSHIP-0592-89>
    NAME: BRIDGESTONE SPORTS TAIWAN CO
    LOCATION: KAOHSIUNG (UNKNOWN)
    TYPE: COMPANY
<ENTITY_RELATIONSHIP-0592-90> :=
    ENTITY1: <ENTITY-0592-87>
            <ENTITY-0592-88>
    ENTITY2: <ENTITY-0592-89>
    REL OF ENTITY2 TO ENTITY1: CHILD
<ENTITY-0592-87> :=
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-90>
    NAME: BRIDGESTONE SPORTS CO
    ALIASES: "BRIDGESTONE SPORTS"
    TYPE: COMPANY
<ENTITY-0592-88> :=
    NAME: TRADING HOUSE
    NATIONALITY: JAPAN (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-90>
<OWNERSHIP-0592-89> :=
    TOTAL-CAPITALIZATION: 20000000 TWD
    OWNED: <ENTITY-0592-89>
<ACTIVITY-0592-89> :=
    ACTIVITY-SITE: (TAIWAN (COUNTRY) -)
    INDUSTRY: <INDUSTRY-0592-90>
<INDUSTRY-0592-90> :=
    INDUSTRY-TYPE: PRODUCTION
    PRODUCT/SERVICE: (39 "GOLF CLUBS")
```

Figure 7 (cont.): PAKTUS Template Fills for the Sample Document

SYSTEM PERFORMANCE

Figure 8 summarizes PRC's scores for MUC-5. The MUC-5 version of the system was incomplete at the time of the final testing. All modules were functional, but many task-specific details were missing. The system was ready for testing on the development corpus only two weeks prior to the final test. Performance was improving rapidly – about one point per day. The main limiting factors for PRC were time and availability of people for development. We directed most of our energy toward the basic engineering, such as generating the template formats, which left little time to address the task-specific linguistic requirements. This resulted in severe undergeneration, which accounted for most of the errors (73 percent undergeneration versus 83 percent overall error rate).

Development Effort

Figure 9 enumerates our activities and level of effort in connection with the MUC-5 task, as well as parallel non-Tipster-specific activities in developing our system. Our total development effort in customizing our system for the MUC-5 testing was 2.1 months, with another month for testing, file management, and other non-developmental activities.

SLOT	POS	ACT	ERR	UND	OVG	SUB
<template> subtotals	348	237	59	48	23	9
<tie-up-relationship> subtotals	1806	992	78	63	33	29
<entity> subtotals	4146	1892	78	66	25	30
<entity-relationship> subtotals	2015	968	82	66	30	39
<activity> subtotals	1112	141	94	92	40	17
<industry> subtotals	1185	336	92	82	37	48
<facility> subtotals	340	7	98	98	0	21
<person> subtotals	372	2	100	100	100	*
<ownership> subtotals	526	43	96	93	14	40
<revenue> subtotals	57	0	100	100	*	*
<time> subtotals	153	1	99	99	0	0
ALL OBJECTS	12060	4619	83	73	29	31
MATCHED ONLY	4330	3151	51	34	9	21

	RECALL	PRECISION		P&R	2P&R	P&2R
ALL OBJECTS	19	49	F-MEASURES	27.06	36.95	21.34
MATCHED ONLY	52	72				
TEXT FILTERING	67	91				

Figure 8: PRC Score Summary

– Non-linguistic engineering	1.1 months
– Task-specific linguistics with Tipster data	1.0
– Documentation, publication prep.	0.3
– Testing, scoring	0.5
– File management, communications	0.2
– Non-Tipster-specific system development	1.9
TOTAL	5.0 months

Figure 9. Breakdown of Development Effort

More than half of the development effort involved non-linguistic engineering of the system for MUC-5 requirements, such as the output template format generation. We also needed to convert the extraction components of the system to accommodate the object oriented nature of the MUC-5 templates. This left us with one month of effort for task-specific linguistic development. The specific changes and additions to the PAKTUS knowledge bases for MUC-5 are enumerated in Figure 10. In parallel with the MUC-5 activity, we devoted 1.9 months of effort to generic development of our system, which may have had some impact on MUC-5 performance.

Limiting Factors

The two areas that could significantly improve performance with modest effort are the definition of task-specific semantic mappings and the output generator. These are highlighted in Figure 11. Very little was done here, however, due to limited time and resources. Only 147 of the semantic mappings were defined. We estimate that about 1,000 would be needed to map the generic linguistic data extracted by PAKTUS into the task-specific representations. That would have required about another month of effort. Within the output generator module, which produces the final output template fills, many functions were incomplete or entirely absent. These are not difficult to implement, but do require time and effort, which were not available.

Knowledge Type	Core System	New/ Mod for MUC-5
Words (Stems)	14,816	387
Tokens	18,928	811
Compounds	343	110
Idioms	88	5
Verb categories	16	0
Nominal categories	404	0
Grammar Arcs	273	9
Grammar States	90	5
Concepts	386	1
Semantic Mappings	0	147
Domain Template	0	11

Figure 10. Additions/Modifications to PAKTUS Knowledge Bases for MUC-5

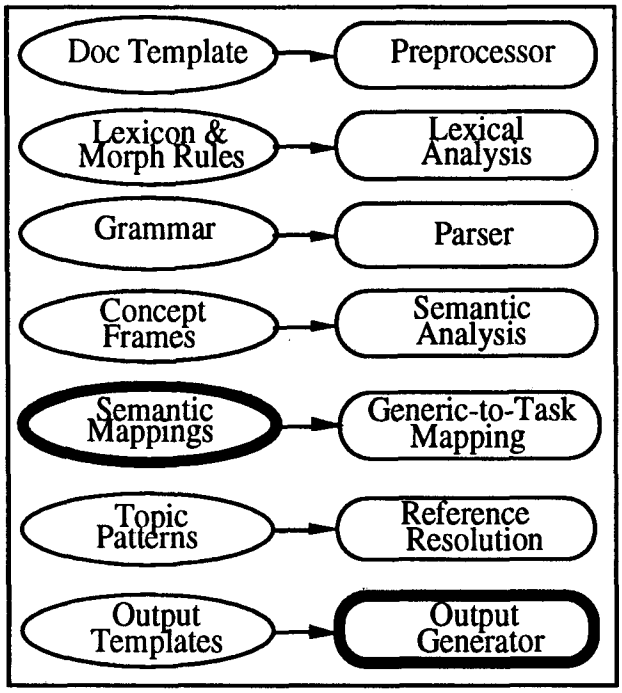

Figure 11: What Would Improve Performance Quickly?

System Training

The PAKTUS modules were trained on varying parts of the MUC-5 corpus. The preprocessing and lexical analysis modules were trained from concordances based on about 1,000 documents. This included the analysis of corporation names for bracketing. The syntactic and semantic analysis modules were largely unchanged, as noted above. The little tailoring that was done was based on a subset of the 86 documents in the dry run, part 1 set. None of this training involved analysis of any complete text, since these modules operate only at the sentence level or below. The careful analysis required for discourse analysis, including coreference resolution, was performed on only two documents: numbers 0099 and 0102. These were selected because 0099 contained multiple tie up relationships, and the other contained a single tie up relationship with

some complex coreference phenomena. This combination seemed to maximize coverage of the phenomena the system had to deal with, within our very limited resource constraints.

Reusability of the System

Almost all of PAKTUS is generic and can be applied to other applications. All of its processes are at least partly generic, as illustrated in Figure 12. They operate on a set of object-oriented knowledge bases, some of which are generic (common English grammar, lexicon, and concept frames) and some of which are task-specific (input and output templates, semantic mappings, and topic patterns). Even within the task-specific knowledge bases, however, the representation schemes are generic, and we have tools that facilitate building them.

The primary tasks in applying PAKTUS to a new domain or improving its performance in an existing domain, are semantic mapping specification, and output generator function development both of which are relatively easy (compared to changing the grammar, for example).

Two other tasks that must be done, but only once for each new domain, are to specify the input document formats and to identify the output specifications. These are template-driven in PAKTUS. For MUC-5 we converted the BNF specifications supplied by the Government to template format, which is quite simple. We then added a function for each template slot to gather information from our generic discourse data structures. Additional information was included regarding output formats, default fills, etc. These templates are also used by a tool for building semantic mappings.

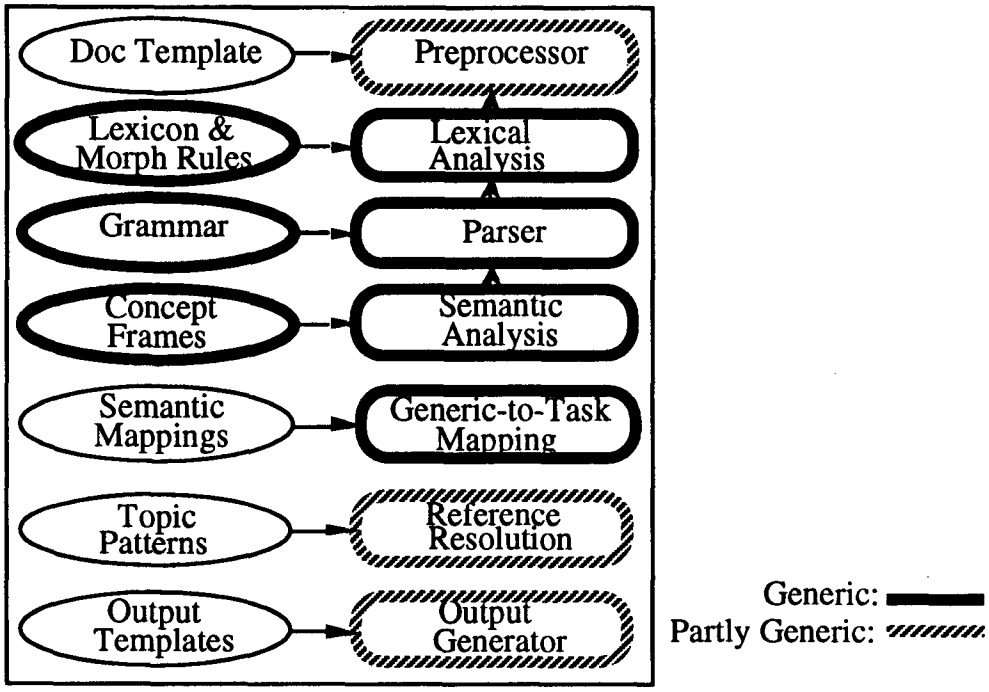

Figure 12: Generic, Reusable Components

Lessons Learned from MUC-5

We confirmed our belief that PAKTUS is robust and adaptable. The more complex components (syntactic, semantic, and discourse analysis modules) are stable and competent enough to apply the system to different domains and produce useful results, by adding domain-specific knowledge (lexicon and semantic mappings). We were once again pleased to learn that it

was not necessary to manually analyze much of the corpus in detail. This was done for only two documents for MUC-5. The full development corpus was used only to customize the preprocessing and lexical analysis components.

A task as complex as MUC-5 requires substantial investment in non-linguistic engineering before the linguistic capabilities of a system can be applied. This detracts from linguistic development that might otherwise have been done, and hides much of the linguistic competence of the system if the engineering is incomplete, as in our case (e.g., correct information was clearly obtained, as in Figure 6, but not reported due to an incomplete output function). We recognize the need for such engineering if useful applications are to be achieved, but hope that this process is standardized quickly so that it does not need to be completely reimplemented for each new application.

REFERENCE

[1] Hobbs, J. "The Generic Information Extraction System", this volume.