STERLING SOFTWARE:

AN NLTOOLSET-BASED SYSTEM FOR MUC-6

Richard Lee

Sterling Software, ITD

1650 Tysons Blvd, #800, McLean VA 22102

Email: rlee@mclean.sterling.com

INTRODUCTION

For a little over two years, Sterling Software ITD has been developing the Automatic Templating System (ATS) [1] for automatically extracting entity and event data in the counter-narcotics domain from military messages. This system, part of the Counter Drug Intelligence System (CDIS), was built around the NLToolset [2], which was originally developed by GE and is now being developed and supported by Lockheed-Martin. Early results showed that the system was performing better than the human analysts in all aspects.

ATS was in its final delivery phase at the same time as our MUC-6 development. We elected to participate despite this conflict, but it did limit us to 4 person-weeks on MUC-6, forcing us to scale back from our original plans and only participate in the NE and TE tasks. The results were more than gratifying.

SYSTEM DESIGN

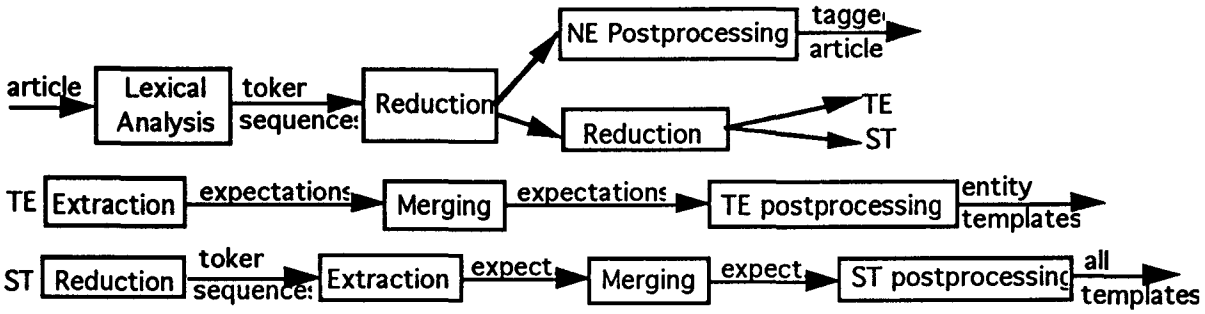

Figure 1: System Design

Our MUC-6 system (Figure 1) consists of 5 major components, applied in sequence: Lexical Analysis, Reduction, Extraction, Merging, Postprocessing. It was designed to share as much of the processing sequence between tasks as possible. The processing for NE followed the identical sequence of steps (Lexical Analysis, and Reduction) as was followed for the TE and ST tasks, then diverged to its own Postprocessing component to write the NE file. The Reduction steps taken to identify portions of text for marking in NE also filled the slots with the appropriate text for the TE task. The processing specific to ST diverged after all the phrase-level Reductions for NE and TE had been performed.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 1995		**00-00-1995 to 00-00-1995**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Sterling Software: An NLToolset-based System for MUC-6	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) **Sterling Software, ITD,1650 Tysons Blvd, #800,McLean,VA,22102**	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
Proceedings of the Sixth Message Understanding Conference (MUC-6), 6-8 Nov 1995, Columbia, MD. Sponsored by the Defense Advanced Research Projects Agency.

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**	**Same as Report (SAR)**	**13**	

The heart of the system is a sophisticated pattern-matcher, which is used repeatedly in the course of processing to identify text for Reduction or Extraction. While the NLToolset also provides a parser, after some initial development we abandoned it on ATS, and did not use it on MUC-6.

Lexical Analysis

The Lexical Analysis component has several subcomponents. First, a tokenizer converts the input string for the entire article into a sequence of tokens. We modified the NLToolset-supplied tokenizer to try to prevent it from reordering or dropping text in ways that made it difficult to map back to the original text when writing the NE output file; we also modified it to preserve upper- vs lower-case information.

The second step in Lexical Analysis is the actual lexicon lookup, which attaches information from the lexicon to the tokens. This includes morphological analysis, which was useful primarily for determining the root form of nationalities, such as "Canadian" -> CANADA. It also includes finding multi-token lexicon entries, such as "New York" and "Coca-Cola". Since we weren't using the parser, the part-of-speech obtained by a lexical lookup was of interest mainly if it was something like city-name or org-name; we did also try to prevent the inappropriate inclusion of verbs, prepositions, etc in names, with mixed results.

The third step in Lexical Analysis is the insertion of special marker tokens to indicate capitalized words. This was needed to be able to usethat information in name recognition, since there did not appear to be any good way to get the pattern matcher to use the capitalization information contained in the original tokens.

Finally, Lexical Analysis splits the token sequence into sentences, including one each for headline, dateline, and date.

Reduction

The Reduction components each consist of one or more stages of applying the NLToolset's pattern matcher to phrases. Any phrase matched is "reduced", usually but not always to a single multi-token, or "mtoken". In each stage, all the patterns appropriate to that stage are tried on each sentence in turn.

The very first reduction stage is a "junk" reduction to delete tables so they are not seen by subsequent reduction stages.

Each subsequent reduction has two useful side-effects: 1) identifying which tokens form the heart of the reduction and therefore should be marked for the NE task, and 2) filling the slots of the mtokens with appropriate pieces of the text that was reduced, for the TE task. Note that these two purposes often conflict -- for example, city, state references and date ranges were supposed to have pieces marked separately, but were reduced to single mtokens with one set of slot fillers. This called for some careful engineering.

The applications of reduction patterns are done in sequence rather than all at once for a number of reasons: First, some references to a person, organization, or location may not be recognizable by themselves, but other references to the same thing may be easier to spot. Therefore, every new thing reduced is added to a temporary lexicon, and another reduction step is applied to look for other references (with certain allowed variations) to those same things; for example, relatively easy-to-recognize references to "Mr. Jones" or "Robert L. James" would enable later recognition of the more problematic "Barnaby Jones" and "James". And when adding to this lexicon, appropriate variations in an (organization) name are included so that they would be recognized if they occured; for example:

Name	Possible variations
"Paramount Pictures Corp."	"Paramount" "Paramount Pictures"
"New York Post"	"Post"
"Kidder , Peabody & Co."	"Kidder" "Kidder Peabody"
"National Labor Relations Board"	"NLRB"

When such a "secondary" organization reference is reduced, the text is put in the org_alias slot; the full form is pulled from the lexicon and put in the org_name slot to ensure proper merging (see below) of the two referents.

Second, the results of reductions can be used to provide additional context for later reductions; for example, person reduction is done after organization, so a reduced organization can help the pattern matcher recognize a person, as in the token sequence [ARTIE MCDONALD , *ORG* 'S PRESIDENT], where *ORG* is the mtoken produced by the earlier reduction. A reduction can also involve multiple previously-reduced mtokens, filling the slots of one with information from another; for example, the reduction of the token sequence [*ORG* , A *LOC* - BASED MANUFACTURER] includes filling the org_descriptor, org_locale, and org_country slots of *ORG* with the descriptive phrase and the information from *LOC*.

Extraction

An Extraction component uses the results of a pattern match to generate an "expectation" and fill its slots with pieces of the text matched. For ST, a typical expectation represents an event, with the person, organization, date, etc mtokens in the clause that was matched being used to fill its slots. For TE, each expectation is a trivial one containing one person or organization.

Merging

The NLToolset provides a merging tool, which merges expectations of the same type (person, organization, etc) as long as the fillers of their corresponding slots do not conflict; a conflict occurs if both have a filler, the fillers are different, and the slot is not allowed to have multiple fillers. Obviously, the org_alias and org_descriptor slots were allowed to have multiple fillers and org_name was not.

During reduction, our system actually splits a person's name across slots called given_name, family_name, and suffix_name, so that the expectations for, say, "Harry L. James, Jr." and "Mr. James" would be merged. It also carefully fills slots such as org_type and a few others added just for this purpose so as to prevent improper merges; for example, it reduces the token sequence [THE *ORG* UNIT] to two *ORG* mtokens, one old and one new, with slots filled so that they could not merge with each other.

Initially, we relied on this merging tool to bring together separated org names and descriptors, such as "NEC Corp. ... the giant Japanese computer manufacturer". We soon found, however, that even with careful use of slot fillers to prevent descriptors for commercial organizations from merging with, say, the name of a government organization or a library, too many merges were incorrect. We therefore devised a separate stack mechanism which keeps track of the org mtokens for each sentence; when an org descriptor is reduced in the final TE reduction stage, the stack is searched starting at the current sentence, to find the closest suitable referent that precedes the descriptor, and to add the descriptor text to the mtoken for that referent. This approach worked quite well.

Postprocessing

For the NE task, the postprocessing step consists of traversing the token sequences in parallel with the original text, writing the original text and inserting markers as the reduction results attached to each token indicated. We had to go back to original text to include those portions of the article header which were not processed, and to recover from cases where the tokenizer had dropped characters despite our modifications.

For the TE task, the postprocessing step consists of traversing the list of expectations and writing a template for each, performing final clean-ups like removing duplicate aliases, combining the person_name pieces, skipping slots used only to control merging, etc.

KNOWLEDGE ENGINEERING

The bulk of the time spent in knowledge engineering was spent developing the patterns for all the Reduction and Extraction stages. These patterns were devised to take advantage of all the local contextual clues we could come up with, including upper- vs lower-case information and descriptive appositives. Our results show that this approach works well; and the modularity of the patterns makes it easy to add coverage as we discover additional clues (such as those we discuss in the walkthrough with respect to organizations).

The reliance on case information meant that headlines were a bit of a problem; despite giving them somewhat special treatment, our error rate was higher there than elsewhere:

*** DOCUMENT SECTION SCORES ***

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
HL	136	142	119	0	7	16	10	0	88	84	7	11	22	6
DD	60	60	60	0	0	0	0	0	100	100	0	0	0	0
DL	52	52	52	0	0	0	0	0	100	100	0	0	0	0
TXT	2046	2024	1889	0	47	88	110	0	92	93	5	4	11	2

There was some lexicon work, as well. This included entries for all the countries, with alternate phrases (such as "West Germany" for "Federal Republic of Germany") and irregular derivations (such as "Dutch" for "Netherlands"), and entries for major cities and geographical regions, with their country information included. For organizations, we limited it to a few dozen major ones that have no reliable internal clues and often occur without any contextual clues (such as "White House", "Fannie Mae", "Big Board", "Coca-Cola" and "Coke", "Macy's", "Exxon", etc).

WALKTHROUGH

NE: 9402240133 key: 9402240133-1 response: 9402240133-1

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
<enamex>	69	68	63	0	0	5	6	0	91	93	9	7	15	0
type	69	68	53	0	10	5	6	0	77	78	9	7	28	16
text	69	68	60	0	3	5	6	0	87	88	9	7	19	5
<timex>	6	9	6	0	0	3	0	0	100	67	0	33	33	0
type	6	9	6	0	0	3	0	0	100	67	0	33	33	0
text	6	9	6	0	0	3	0	0	100	67	0	33	33	0
<numex>	6	7	6	0	0	1	0	0	100	86	0	14	14	0
type	6	7	6	0	0	1	0	0	100	86	0	14	14	0
text	6	7	6	0	0	1	0	0	100	86	0	14	14	0
TOTAL	162	168	137	0	13	18	12	0	84	82	7	11	24	9

TE: 9402240133 key: 9402240133 response: 9402240133

SLOT	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	ERR	SUB
organization	10	9	8	0	1	1	0	0	80	89	10	0	20	11
name	10	8	7	0	1	2	0	0	70	88	20	0	30	13
alias	3	1	1	0	0	2	0	6	33	100	67	0	67	0
descriptor	3	2	1	0	1	1	0	6	33	50	33	0	67	50
type	10	8	8	0	0	2	0	0	80	100	20	0	20	0
locale	2	1	1	0	0	1	0	7	50	100	50	0	50	0
country	2	1	1	0	0	1	0	7	50	100	50	0	50	0
person	6	10	6	0	0	0	4	0	100	60	0	40	40	0
name	6	10	6	0	0	0	4	0	100	60	0	40	40	0
alias	3	4	2	0	0	1	2	4	67	50	33	50	60	0
title	2	3	2	0	0	0	1	4	100	67	0	33	33	0
TOTAL	41	38	29	0	2	10	7	34	71	76	24	18	40	6

Table 1: Results on the walkthrough article

The results on the walkthrough article (see Table 1) compared to our overall results show that this was indeed a relatively difficult article. They show three issues worth discussing.

First, we had low precision on timex. Two out of the three "spurious" dates are due to our apparently mistaken belief that "yesterday" and "tomorrow" were supposed to be marked. This knowledge engineering error led to the worst recall or precision number on our overall NE results, a precision on timex of 84; avoiding that error would have raised it to 94.

Second, recall and precision on organizations was a bit low. The system missed both "Fallon McElligott" and "McCann-Erickson". On the former, a phrase like "ad agency Fallon McElligott" would have caused it to be found, but the actual phrase "other ad agencies, such as Fallon McElligott" did not. On the latter, not having a pattern to cover things like "chief executive officer of McCann-Erickson" was an omission on our part.

Other organization errors were: getting "New York Times", which in this article is incorrect; missing the two descriptors for "Ammirati & Puris" and the locale for "Coca-Cola". The locale error points out another major cause of poor results -- a next-to-last-minute change in the final TE pattern for picking up combination of organization name plus location and/or descriptor, inadequately tested, led to inadvertantly dropping coverage of the most basic of combinations: [*ORG* $lprep *LOC*], where $lprep is a macro for: "one of ',' 'in' 'of'". This unfortunate error had the following effect on total locale slot score on TE:

	POS	ACT	COR	REC	PRE
actual	110	59	42	38	71
corrected	110	76	59	54	71

Third, problems with persons. The system decided "McCann" was a person, based on "the McCann family"; since it did not recognize "McCann-Erickson" as a company, every reference to "McCann" was therefore marked as a person. Due to inadequate restrictions on our use of capitalization, the system also decided "While McCann" and "One McCann" were distinct persons. It decided that "John J. Dooner, Jr." and "John Dooner" were distinct persons; the "Jr." would not have caused it to make that decision, but the "J." did.

Now, the walkthrough.

(*SO-HL* *CAP* MARKETING *AMPERSAND* *CAP* MEDIA *DASHES* *CAP* ADVERTISING *COLON* *AT* *CAP* JOHN *CAP* DOONER *CAP* WILL *CAP* SUCCEED *CAP* JAMES *AT* *CAP* AT *CAP* HELM OF *CAP* MCCANN *HYPHEN* *CAP* ERICKSON *AT* *DASHES* *AT* *CAP* BY *CAP* KEVIN *CAP* GOLDMAN *EO-HL*)

(*SO-DD* |02| *SLASH* |24| *SLASH* |94| *EO-DD*)

(*SO-P* *CAP* ONE OF THE MANY DIFFERENCES BETWEEN *CAP* ROBERT *CAP* ABBREV_L *CAP* JAMES *COMMA* CHAIRMAN AND CHIEF-EXECUTIVE-OFFICER OF *CAP* MCCANN *HYPHEN* *CAP* ERICKSON *COMMA* AND *CAP* JOHN *CAP* ABBREV_J *CAP* DOONER *CAP* ABBREV_JR *COMMA* THE AGENCY *APOSTROPHE-S* PRESIDENT AND CHIEF-OPERATING-OFFICER *COMMA* IS QUITE TELLING *COLON* *CAP* ABBREV_MR *CAP* JAMES ENJOYS SAILBOATING *COMMA* WHILE *CAP* ABBREV_MR *CAP* DOONER OWNS A POWERBOAT *PERIOD*)

...

(*CAP* HOWEVER *COMMA* ODDS OF THAT HAPPENING ARE SLIM SINCE WORD FROM *CAP* COCA-COLA HEADQUARTERS IN *CAP* ATLANTA IS THAT *CAP* CAA AND OTHER AD AGENCIES *COMMA* SUCH AS *CAP* FALLON *CAP* MCELLIGOTT *COMMA* WILL CONTINUE TO HANDLE *CAP* COCA-COLA ADVERTISING *PERIOD*)

...

(*DOUBLEQUOTE* *EO-P* *SO-P* *CAP* ABBREV_MR *CAP* DOONER MET WITH *CAP* MARTIN *CAP* PURIS *COMMA* PRESIDENT AND CHIEF-EXECUTIVE-OFFICER OF *CAP* AMMIRATI *AMPERSAND* *CAP* PURIS *COMMA* ABOUT *CAP* MCCANN *APOSTROPHE-S* ACQUIRING THE AGENCY WITH BILLINGS OF *DOLLAR* |400| MILLION *COMMA* BUT NOTHING HAS MATERIALIZED *PERIOD*)

...

Figure 2: After Lexical Analysis

After the Lexical Analysis, the input string has been converted into a list of 52 sentences, each sentence containing a list of tokens; this list includes *CAP* tokens inserted in front of every capitalized token. Attached to each token is the result of the lexical lookup.

Note that at this point lexical lookup has replaced the surface representation of "Coke" and "CEO" with their "canonical" forms. Every token contains its original string, so we can still recover it for use in filling slots.

The lookup on "Atlanta" has provided the information that it is a city and that its country is the US.

(*SO-HL* *CAP* MARKETING *AMPERSAND* *CAP* MEDIA *DASHES* *CAP* ADVERTISING *COLON* *AT* *IND* *CAP* WILL *CAP* SUCCEED *IND* *AT* *CAP* AT *CAP* HELM OF *IND* *HYPHEN* *CAP* ERICKSON *AT* *DASHES* *AT* *CAP* BY *IND* *EO-HL*)

(*SO-DD* *DATE* *EO-DD*)

(*SO-P* *CAP* ONE OF THE MANY DIFFERENCES BETWEEN *IND* *COMMA* CHAIRMAN AND CHIEF-EXECUTIVE-OFFICER OF *IND* *HYPHEN* *CAP* ERICKSON *COMMA* AND *IND* *COMMA* THE AGENCY *APOSTROPHE-S* PRESIDENT AND CHIEF-OPERATING-OFFICER *COMMA* IS QUITE TELLING *COLON* *CAP* *IND* ENJOYS SAILBOATING *COMMA* WHILE *CAP* *IND* OWNS A POWERBOAT *PERIOD*)

...

(*CAP* HOWEVER *COMMA* ODDS OF THAT HAPPENING ARE SLIM SINCE WORD FROM *ORG* HEADQUARTERS IN *LOC* IS THAT *CAP* CAA AND OTHER AD AGENCIES *COMMA* SUCH AS *CAP* FALLON *CAP* MCELLIGOTT *COMMA* WILL CONTINUE TO HANDLE *ORG* ADVERTISING *PERIOD*)

...

(*DOUBLEQUOTE* *EO-P* *SO-P* *CAP* *IND* MET WITH *IND* *COMMA* PRESIDENT AND CHIEF-EXECUTIVE-OFFICER OF *ORG* *COMMA* ABOUT *IND* *APOSTROPHE-S* ACQUIRING THE AGENCY WITH BILLINGS OF *MONEY* *COMMA* BUT NOTHING HAS MATERIALIZED *PERIOD*)

...

Figure 3: After Entity Reductions

The initial Reduction stages take care of money, percent, date, time, and location, then "secondary" references to location. The only things worth noting here are the "yesterday" errors already discussed, that the system decided "60 pounds" was a reference to money, and that the information in the lexical entry for "Atlanta" was used to fill the slots of the *LOC* mtoken.

The next Reduction stages take care of "primary" then "secondary" references to organizations.

The primary stage picks up "Interpublic Group", "PaineWebber", "Coca-Cola", "Coke", "Creative Artists Agency", "WPP Group", "Ammirati & Puris", "New York Yacht Club" and "New York Times". It misses "Fallon McBride" and "McCann-Erickson" for reasons already noted. The only reason it gets "PaineWebber", "Coca-Cola", and "Coke" is because they are in the lexicon; the others are all picked up by match various patterns.

In this article, the only secondary reference is "CAA" as a reference to "Creative Artists Agency". While the system does manufacture acronyms as potential secondary references when certainpatterns match, the pattern which enabled it to determine that "Creative Artists Agency" was a commercial organization was unfortunately not one of them.

The next Reduction stages take care of "primary" then "secondary" references to persons.

The primary stage picks up "James", "John Dooner", "Kevin Goldman", "Robert L. James", "John J. Dooner, Jr.", "Mr. James", "Mr. Dooner", "Alan Gottesman", "Peter Kim", "Walter Thompson", "Martin Puris", and (alas) "McCann". These are found on the strength of titles like "Mr." and "Sen.", known first names, and contextual clues such as known occupations like "president", "analyst", etc. "James" in the headline is found because it follows "succeed"; "McCann" is found because of "McCann family".

The secondary stage picks up all remaining references to "McCann". Since "McCann-Erickson" was not recognized as an organization, all those occurrences are picked up, too. And since we failed to make adverbs off-limits as new first names in this stage, it decides that "While McCann" and "One McCann" (note the capitalization) are distinct persons.

<DOC>
<DOCID> wsj94_026.0231 </DOCID>
<DOCNO> 940224-0133. </DOCNO>
<HL> Marketing & Media -- Advertising:
@ <ENAMEX TYPE="PERSON">John Dooner</ENAMEX> Will Succeed
<ENAMEX TYPE="PERSON">James</ENAMEX>
@ At Helm of <ENAMEX TYPE="PERSON">McCann</ENAMEX>-Erickson
@ ----
@ By <ENAMEX TYPE="PERSON">Kevin Goldman</ENAMEX> </HL>
<DD> <TIMEX TYPE="DATE">02/24/94</TIMEX> </DD>
<SO> WALL STREET JOURNAL (J), PAGE B8 </SO>
<CO> IPG K </CO>
<IN> ADVERTISING (ADV), ALL ENTERTAINMENT & LEISURE (ENT),
FOOD PRODUCTS (FOD), FOOD PRODUCERS, EXCLUDING FISHING (OFP),
RECREATIONAL PRODUCTS & SERVICES (REC), TOYS (TMF) </IN>
<TXT>
<p>
One of the many differences between <ENAMEX TYPE="PERSON">Robert L. James</ENAMEX>, chairman and chief executive officer of <ENAMEX TYPE="PERSON">McCann</ENAMEX>-Erickson, and <ENAMEX TYPE="PERSON">John J. Dooner Jr.</ENAMEX>, the agency's president and chief operating officer, is quite telling: Mr. <ENAMEX TYPE="PERSON">James</ENAMEX> enjoys sailboating, while
Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> owns a powerboat.

...

However, odds of that happening are slim since word from
<ENAMEX TYPE="ORGANIZATION">Coke</ENAMEX> headquarters in
<ENAMEX TYPE="LOCATION">Atlanta</ENAMEX> is that CAA and other ad agencies, such as Fallon McElligott, will continue to handle <ENAMEX TYPE="ORGANIZATION">Coke</ENAMEX> advertising.

...

Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> met with
<ENAMEX TYPE="PERSON">Martin Puris</ENAMEX>, president and chief executive officer of
<ENAMEX TYPE="ORGANIZATION">Ammirati & Puris</ENAMEX>, about
<ENAMEX TYPE="PERSON">McCann</ENAMEX>'s acquiring the agency with billings of
<NUMEX TYPE="MONEY">$400 million</NUMEX>, but nothing has materialized.

...

Figure 4: NE Results

Now, NE and TE processing diverge. For NE, the system uses the original text of the article to write a copy. It traverses the token sequences in parallel with the original text, using the fact that each token contains information on all the reductions it was involved in to determine where to insert begin and end brackets. It only pays attention to the final reduction except in the case of locations inside money, where brackets are inserted for both.

Before	After
Person: Family_Name: "James" Per_Alias: "James"	Person: Per_Title: "Mr." Given_Name: "Robert L." Family_Name: "James" Per_Alias: "James"
Person: Given_Name: "Robert L." Family_Name: "James"	
Person: Per_Title: "Mr." Family_Name: "James" Per_Alias: "James"	
Person: Given_Name: "While" Family_Name: "McCann"	Person: Given_Name: "While" Family_Name: "McCann" Per_Alias: "McCann"
Person: Given_Name: "One" Family_Name: "McCann"	Person: Given_Name: "One" Family_Name: "McCann"
Person: Family_Name: "McCann" Per_Alias: "McCann"	
Organization: Org_Name: "Coca-Cola" Org_Alias: "Coca-Cola" Known: "Yes"	Organization: Org_Name: "Coca-Cola" Org_Alias: "Coca-Cola" "Coke" Known: "Yes"
Organization: Org_Name: "Coca-Cola" Org_Alias: "Coke" Known: "Yes"	

Figure 5: Merging

For TE, there is one final Reduction stage to take care of organization descriptors and locations. Here, the system finds descriptors "the big Hollywood talent agency" and "a hot agency", but not "a quality operation" and "the agency with billings of $400 million". The former omission was deliberate, due to too many spurious matches when it was included; the latter was a construct we did not think to include. In

cases where the descriptor is an appositive, the referenced organization is included in the pattern match; otherwise, if the appositive is a definite reference, the stack of organization references is searched for the putative antecedant. In either case, the descriptor and locale information (if any) is inserted into slots of the organization mtoken. In retrospect, including indefinite references that are not appositives appears to have been the wrong thing to do.

Then there is the trivial Extraction step which turns the organization and person mtokens into "expectations". This is followed by the Merging step which merges expectations together wherever possible. This includes merging the expectations for "James", "Robert L. James", "Mr. James" (several occurences); "Coca-Cola", "Coke"; etc.

Before	After
Person: Per_Title: "Mr." Given_Name: "Robert L." Family_Name: "James" Per_Alias: "James"	Person: Per_Title: "Mr." Per_Name: "Robert L. James" Per_Alias: "James"
Organization: Org_Name: "Coca-Cola" Org_Alias: "Coca-Cola" "Coke" Known: "Yes"	Organization: Org_Name: "Coca-Cola" Org_Alias: "Coke"

Figure 6: Postprocess Slot Adjustments

Finally, the Postprocessing step writes each expectation to the TE result file, making final adjustments to the slot fillers as needed.

```
<PERSON-9402240133-1> :=
  PER_TITLE:  "Mr."
  PER_NAME:   "John Dooner"
  PER_ALIAS:  "Dooner"

<PERSON-9402240133-2> :=
  PER_TITLE:  "Mr."
  PER_NAME:   "Robert L. James"
  PER_ALIAS:  "James"

<PERSON-9402240133-3> :=
  PER_NAME:   "While McCann"
  PER_ALIAS:  "McCann"

<PERSON-9402240133-4> :=
  PER_NAME:   "Kevin Goldman"

<PERSON-9402240133-5> :=
  PER_TITLE:  "Mr."
  PER_NAME:   "John J. Dooner Jr."
  PER_ALIAS:  "Dooner"

ORGANIZATION-9402240133-1> :=
  ORG_TYPE:   "COMPANY"
  ORG_NAME:   "Interpublic Group"

<PERSON-9402240133-6> :=
  PER_NAME:   "Alan Gottesman"

<ORGANIZATION-9402240133-2> :=
  ORG_TYPE:   "COMPANY"
  ORG_NAME:   "PaineWebber"

<ORGANIZATION-9402240133-3> :=
  ORG_TYPE:   "COMPANY"
  ORG_NAME:   "Coca-Cola"
  ORG_ALIAS:  "Coke"

<ORGANIZATION-9402240133-4> :=
  ORG_DESCRIPTOR: "the big Hollywood talent agency"
  ORG_COUNTRY:    "UNITED STATES"
  ORG_LOCALE:     "Hollywood CITY"
  ORG_TYPE:       "COMPANY"
  ORG_NAME:       "Creative Artists Agency"

<PERSON-9402240133-7> :=
  PER_NAME:   "One McCann"

<PERSON-9402240133-8> :=
  PER_NAME:   "Peter Kim"

<ORGANIZATION-9402240133-5> :=
  ORG_TYPE:   "COMPANY"
  ORG_NAME:   "WPP Group"

<PERSON-9402240133-9> :=
  PER_NAME:   "Walter Thompson"

<ORGANIZATION-9402240133-6> :=
  ORG_DESCRIPTOR: "a hot agency"

<PERSON-9402240133-10> :=
  PER_NAME:   "Martin Puris"

<ORGANIZATION-9402240133-7> :=
  ORG_TYPE:   "COMPANY"
  ORG_NAME:   "Ammirati & Puris"

<ORGANIZATION-9402240133-8> :=
  ORG_TYPE:   "OTHER"
  ORG_NAME:   "New York Yacht Club"

<ORGANIZATION-9402240133-9> :=
  ORG_TYPE:   "COMPANY"
  ORG_NAME:   "New York Times"
```

Figure 7: TE Results

RESULTS AND CONCLUSIONS

NE: * * TOTAL SLOT SCORES * * *

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
<enamex>	942	914	884	0	0	30	58	0	94	97	6	3	9	0
type	942	914	866	0	18	30	58	0	92	95	6	3	11	2
text	942	914	858	0	26	30	58	0	91	94	6	3	12	3
subtotals	1884	1828	1724	0	44	60	116	0	92	94	6	3	11	2
<timex>	112	132	111	0	0	21	1	0	99	84	1	16	16	0
type	112	132	111	0	0	21	1	0	99	84	1	16	16	0
text	112	132	101	0	10	21	1	0	90	76	1	16	24	9
subtotals	224	264	212	0	10	42	2	0	95	80	1	16	20	4
<numex>	93	93	92	0	0	1	1	0	99	99	1	1	2	0
type	93	93	92	0	0	1	1	0	99	99	1	1	2	0
text	93	93	92	0	0	1	1	0	99	99	1	1	2	0
subtotals	186	186	184	0	0	2	2	0	99	99	1	1	2	0
ALL OBJ	2294	2278	2120	0	54	104	120	0	92	93	5	4	12	2
F-MEASURES								92.74	92.93	92.54				

TE: * * * TOTAL SLOT SCORES * * *

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	ERR	SUB
organization	606	574	506	0	21	79	47	0	83	88	13	8	23	4
name	546	512	397	0	72	77	43	21	73	78	14	8	33	15
alias	173	105	84	0	2	87	19	335	49	80	50	18	56	2
descriptor	235	175	89	0	27	119	59	285	38	51	51	34	70	23
type	606	538	472	0	34	100	32	0	78	88	17	6	26	7
locale	115	59	42	0	8	65	9	218	37	71	57	15	66	16
country	116	59	48	0	2	66	9	217	41	81	57	15	62	4
person	496	506	457	0	14	25	35	0	92	90	5	7	14	3
name	496	506	443	0	28	25	35	0	89	88	5	7	17	6
alias	170	166	160	0	0	10	6	264	94	96	6	4	9	0
title	166	177	165	0	0	1	12	265	99	93	1	7	7	0
ALL OBJ	2623	2297	1900	0	173	550	224	1606	72	83	21	10	33	8
F-MEASURES								77.24	80.43	74.28				

Table 2: Overall Scores

The overall results (see Table 2) were obtained in 4 person-weeks of effort, lifting some pattern and code ideas from the ATS, which worked on a very different set of message types, and wasting a few days on the ST task and on filling in date templates. These results show that our semantic-pattern-based approach to entity detection and templating is a very good one, and one which can be brought to bear on a new application quickly.

As we have noted, dramatic improvements in the worst numbers (timex in NE, org locale and country in TE) would have been obtained with very minor changes in the patterns -- literally, a couple hours worth of work. The org locale fix would actually have given us the highest f-measure on that category: 61.3. Despite that "couple hours" estimate, we would have to say that our greatest limiting factor was

time -- time to test more thoroughly and isolate the causes of the biggest problems. Slowness of the system was a problem but not a major one, as it took only a minute or two per article.

After those two improvements, we turn to the problem of org descriptors -- although we had the highest f-measure, it was only 43.6, which shows that there is still room for improvement. Here, the solutions are less obvious. One step to take is to add to the patterns to allow modifier phrases after the head noun in a descriptor noun phrase, such as "the agency with billings of $400 million". More exploration is needed on this, especially in light of the fact that both the recall and precision rates were low.

Another area where we would like to make changes is in the order of reduction stages. For example, the system currently does all person reductions after organization reductions. This meant we had to prevent the secondary organization reduction from matching what are clearly person names (eg: primary "Schecter Group" -/-> secondary "Mr. Schecter"). The solution, clearly, is to apply some of the person patterns before the organization patterns.

Since all the processing occurs without any regard to the types of events discussed in the articles, the system we have developed here is easily portable across domains. If a domain required a different set of template slots than used for MUC-6, the patterns would be unchanged but the reduction code that fills the slots, and the postprocessing code that reports them, would have to be modified slightly.

We have demonstrated, on MUC-6 and on CDIS, that we have an excellent approach to both entity and event extraction on a range of document types. We hope to have the opportunity to continue this work, as funding permits.

REFERENCES

[1] Osterholtz, L. and Lee, R. and McNeilly, C., "The Automated Templating System for Database Update from Unformatted Message Traffic", *1995 ONDCP International Technology Symposium*, Oct. 1995

[2] Jacobs, P.S. and Krupka, G.R. and Rau, L.F., "Lexico-Semantic Pattern Matching as a Companion to Parsing in Text Applications", *Fourth DARPA Speech and Natural Language Workshop*, Feb. 1991