UMASS/HUGHES: DESCRIPTION OF THE CIRCUS SYSTEM USED FOR MUC-5[1]

W. Lehnert, J. McCarthy, S. Soderland, E. Riloff, C. Cardie, J. Peterson, F. Feng
University of Massachusetts
Department of Computer Science
Box 34610
Amherst, MA 01003-4610
lehnert@cs.umass.edu

C. Dolan, S. Goldman
Hughes Research Laboratories
3011 Malibu Canyon Road M/S RL96
Malibu, CA 90265
cpd@aic.hrl.hac.com

INTRODUCTION

The primary goal of our effort is the development of robust and portable language processing capabilities for information extraction applications. The system under evaluation here is based on language processing components that have demonstrated strong performance capabilities in previous evaluations [Lehnert et al. 1992a]. Having demonstrated the general viability of these techniques, we are now concentrating on the practicality of our technology by creating trainable system components to replace hand-coded data and manually-engineered software.

Our general strategy is to automate the construction of domain-specific dictionaries and other language-related resources so that information extraction can be customized for specific applications with a minimal amount of human assistance. We employ a hybrid system architecture that combines selective concept extraction [Lehnert 1991] technologies developed at UMass with trainable classifier technologies developed at Hughes [Dolan et al. 1991]. Our MUC-5 system incorporates seven trainable language components to handle (1) lexical recognition and part-of-speech tagging, (2) knowledge of semantic/syntactic interactions, (3) semantic feature tagging, (4) noun phrase analysis, (5) limited coreference resolution, (6) domain object recognition, and (7) relational link recognition. Our trainable components have been developed so domain experts who have no background in natural language or machine learning can train individual system components in the space of a few hours.

Many critical aspects of a complete information extraction are not appropriate for customization or trainable knowledge acquisition. For example, our system uses low-level text specialists designed to recognize dates, locations, revenue objects, and other common constructions that involve knowledge of conventional language. Resources of this type are portable across domains (although not all domains require all specialists) and should be developed as sharable language resources. The UMass/Hughes focus has been on other aspects of information extraction that can benefit from corpus-based knowledge acquisition. For example, in any given information extraction application, some sentences are more important than others, and within a single sentence some phrases are more important than others. When a dictionary is customized for a specific application, vocabulary coverage can be sensitive to the fact that a lot of words contribute little or no information to the final extraction task: full dictionary coverage is not needed for information extraction applications.

In this paper we will overview our hybrid architecture and trainable system components. We will look at examples taken from our official test runs, discuss the test results obtained in our official and optional test runs, and identify promising opportunities for additional research.

[1] This work was supported by the Department of Defense under Contract No. MDA904-92-C-2390. The views expressed in this paper are those of the authors and should not be interpreted to represent opinions or policies of the United States Government.

TRAINABLE LANGUAGE PROCESSING

Our MUC-5 system relies on two major tools that support automated dictionary construction: (1) OTB, a trainable part-of-speech tagger, and (2) AutoSlog, a dictionary construction tool that operates in conjunction with the CIRCUS sentence analyzer. We trained OTB for EJV on a subset of EJV texts and then again for EME using only EME texts. OTB is notable for the high hit rates it obtains on the basis of relatively little training. We found that OTB attained overall hit rates of 97% after training on only 1009 sentences for EJV. OTB crossed the 97% threshold in EME after only 621 training sentences. Incremental OTB training requires human interaction with a point-and-click interface. Our EJV training was completed after 16 hours with the interface; our EME training required 10 hours.

AutoSlog is a dictionary construction tool that analyzes source texts in conjunction with associated key templates (or text annotations) in order to propose concept node (CN) definitions for CIRCUS [Riloff & Lehnert 1993; Riloff 1993]. A special interface is then used for a manual review of the AutoSlog definitions in order to separate the good ones from the bad ones. Of 3167 AutoSlog CN definitions proposed in response to 1100 EJV key templates, 944 (30%) were retained after manual inspection. For EME, AutoSlog proposed 2952 CN definitions in response to 1000 key templates and 2275 (77%) of these were retained after manual inspection. After generalizing the original definitions with active/passive transformations, verb tense generalizations, and singular/plural generalizations, our final EJV dictionary contained 3017 CN definitions and our final EME dictionary contained 4220 CN definitions. It took 20 hours to manually inspect and filter the full EJV dictionary; the full EME dictionary was completed in 17 hours. The CIRCUS dictionary used in our official run was based exclusively on AutoSlog CN definitions. No hand-coded or manually altered definitions were added to the CN dictionary.

When CIRCUS processes a sentence it can invoke a semantic feature tagger (MayTag) that dynamically assigns features to nouns and noun modifiers. MayTag uses a feature taxonomy based on the semantics of our target templates, and it dynamically assigns context-sensitive tags using a corpus-driven case-based reasoning algorithm [Cardie 93]. MayTag operates as an optional enhancement to CIRCUS sentence analysis. We ran CIRCUS with MayTag for EJV, but did not use it for EME (we'll return to a discussion of this and other domain differences later). MayTag was trained on 174 EJV sentences containing 5591 words (3060 open class words and 2531 closed class words). Our tests indicate that MayTag achieves a 74% hit rate on general semantic features (covering 14 possible tags) and a 75% hit rate on specific semantic features (covering 42 additional tags). Interactive training for MayTag took 14 hours using a text editor.

An important aspect of the MUC-5 task concerns information extraction at the level of noun phrases. Important set fill information is often found in modifiers, such as adjectives and prepositional phrases. Part-of-speech tags help us identify basic noun phrase components, but higher-level processes are needed to determine if a prepositional phrase should be attached, how a conjunction should be scoped, or if a comma should be crossed. Noun phrase recognition is a non-trivial problem at this higher level. To address the more complicated aspects of noun phrase recognition, we use a trainable classifier that attempts to find the best termination point for a relevant noun phrase. This component was trained exclusively on the EJV corpus and then used without alteration for both EJV and EME. Experiments indicate that the noun phrase classifier terminates EJV noun phrases perfectly 87% of the time. 7% of its noun phrases pick up spurious text (they are extended too far), and 6% are truncated (they are not extended or extended far enough). Similar hit rates are found with EME test data: 86% for exact NP recognition, with 6% picking up spurious text and 8% being truncated. The noun phrase classifier was trained on 1350 EJV noun phrases examined in context. It took 14 hours to manually mark these 1350 instances using a text editor.

Before we can go from CIRCUS output to template instantiations, we create intermediate structures called memory tokens. Memory tokens incorporate coreference decisions and structure relevant information to facilitate template generation. Memory tokens record source strings from the original input text, OTB tags, MayTag features, and pointers to concept nodes that extracted individual noun phrases.

Discourse analysis contributes to critical decisions associated with memory tokens. Here we find the greatest challenges to trainable language systems. Thus far, we have implemented one trainable component that contributes to coreference resolution in limited contexts. We isolate compound noun phrases that are syntactically consistent with appositive constructions and pass these NP pairs on to a coreference classifier. Since adjacent NPs may be separated by a comma if they occur in a list or at a clause boundary, it is easy to confuse legitimate appositives with pairings of unrelated (but adjacent) NPs. Appositive recognition is therefore treated as a binary classification problem that can be handled with corpus-driven training. For our official MUC-5 runs we trained a classifier to handle appositive recognition using EJV development texts and then used the resulting classifier for both EJV and EME. Our best test results with this classifier showed an 87% hit rate on EJV appositives. It took 10 hours to manually classify 2276 training instances for the appositive classifier using a training interface.

Our final tool, TTG, is responsible for the creation of template generators that map CIRCUS output into final template instantiations. TTG template generators are responsible for the recognition and creation of domain objects as well as the insertion of relational links between domain objects. TTG is corpus-driven and requires no human intervention during training. Application-specific access methods (pathing functions) must be hand-coded for a new domain but these can be added to TTG in a few days by a knowledgeable technician working with adequate domain documentation. Once these adjustments are in place, TTG uses memory tokens and key templates to train classifiers for template generation. No further human intervention is required to create the template generators, although additional testing, tuning and adjustments are needed for optimal performance.

Our hybrid architecture demonstrates how machine learning capabilities can be utilized to acquire many different kinds of knowledge from a corpus. These same acquisition techniques also make it easy to exploit the resulting knowledge without additional knowledge engineering or sophisticated reasoning. The knowledge we can readily acquire from a corpus of representative texts is limited with respect to reusability, but nevertheless cost-effective in a system development scenario predicated on customized software. The trainable components used for *both* EJV and EME were completed after 101 hours of interactive work by a human-in-the-loop. Moreover, most of our training interfaces can be effectively operated by domain experts: programming knowledge or familiarity with computational linguistics is generally not required.[2] Near the end of this paper will report the results of a system development experiment that supports this claim.

There will always be a need for some amount of manual programming during the system development cycle for a new information extraction application. Even so, significant amounts of system development that used to rely on experienced programmers have been shifted over to trainable language components. The ability to automate knowledge acquisition on the basis of key templates represents a significant redistribution of labor away from skilled knowledge engineers, who need access to domain knowledge, directly to the domain experts themselves. By putting domain experts into the role of the human-in-the-loop we can reduce dependence on software technicians. When significant amounts of system development work is being handled by automated knowledge acquisition and expert-assisted knowledge acquisition, it will become increasingly cost-effective to customize and maintain a variety of information extraction applications. We have only just begun to explore the range of possibilities associated with trainable language processing systems.

The hybrid architecture underlying our official MUC-5 systems was less than six months old at the time of the evaluation, and most of the trainable language components that we utilized were less than a year old. Less than 24 person/months were expended for both of the EJV and EME systems, although this estimate is confounded by the fact that trainable components and their associated interfaces were being designed, implemented, and tested by the same people responsible for our MUC-5 system development. The creation of a trainable system component represents a one-time system development investment that can be applied to subsequent systems at much less overhead.

Figure 1 outlines the basic flow-of-control through the major components of the UMass/Hughes MUC-5 system. Note that most of the trainable components depend only on the texts from the development corpus. The concept node dictionary and the trainable template generator also rely on answer keys during training. In the case of the concept node dictionary, we have been able to drive our dictionary construction process on the basis of annotated texts created by using a point-and-click text marking interface. So the substantial overhead associated with creating a large collection of key templates is not needed to support automated dictionary construction. However, we do not see how to support trainable template generation without a set of key templates, so this one trainable component requires a significant investment with respect to labor.

[2] Some technical background is needed to train OTB. Knowledge of our part-of-speech tags is needed for that interface.

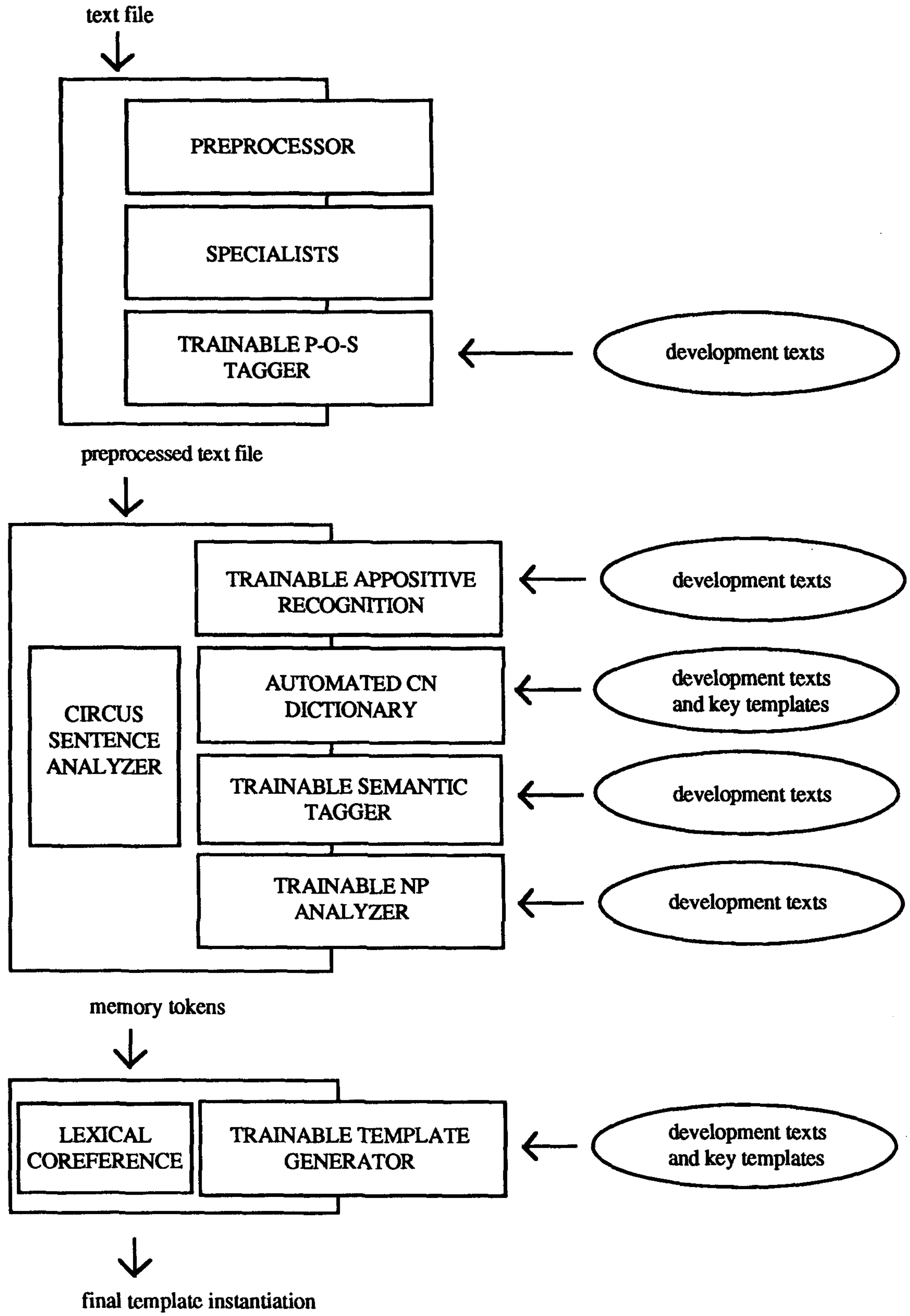

Figure 1: System Architecture

THE OFFICIAL TEST RUNS

Our official test runs were conducted on four DECstations running Allegro Common Lisp. The runs went smoothly in EME but we encountered one fatal error in one of the EJV test sets. Portions of our official EJV and EME score reports are shown in figures 2 and 3.

	ERR	SUB	REC	PRE	UND	OVG
all objects	77	25	26	54	65	28
matched only	48	17	55	76	34	8
text filtering			65	94	35	6

	P&R	2P&R	P&2R
F-measures	35.18	44.39	29.13

Figure 2: Official EJV Score Report Summaries

	ERR	SUB	REC	PRE	UND	OVG
all objects	77	24	31	39	59	48
matched only	54	15	60	59	30	31
text filtering			85	76	15	23

	P&R	2P&R	P&2R
F-measures	34.84	37.42	32.59

Figure 3: Official EME Score Report Summaries

THE OPTIONAL TEST RUNS

We ran optional tests to see what sort of recall/precision trade-offs were available from the system. Since the template generator is a set of classifiers, and each classifier outputs a certainty associated with a hypothesized template fragment, we have many parameters that can be manipulated. Raising the threshold on the certainty for a hypothesis will, in most cases, increase precision and reduce recall. In the experiments reported here, we have varied the parameters over broad classes of discrimination trees. There are three important classes of decision tree: (1) trees that filter the creation of objects based on string fills, (2) trees that filter the creation of objects based on set fills, and (3) trees that hypothesize relations among objects. An example of the first class is the tree that filters the CIRCUS output for entity names in the EJV domain. An example of the second class is the tree that filters possible lithography objects based on evidence of the type of lithography process. The trees that hypothesize TIE_UP_RELATIONSHIP's and ME_CAPABILITY's are examples of the third class.

For these experiments we have varied the certainty thresholds for all trees of a given class. Figure 4 shows the trade-off achieved for EME.

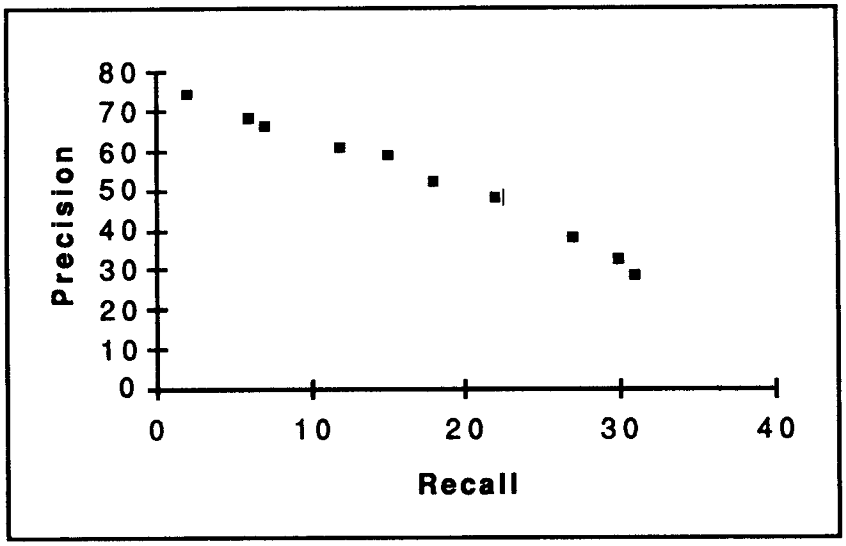

Figure 4: Trade-off curve for EME

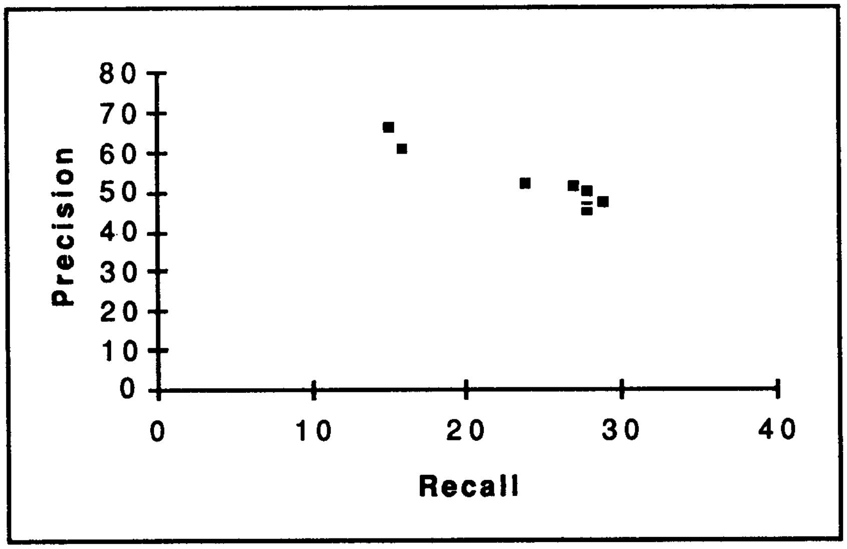

Figure 5: Trade-off curve for EJV

This trade-off curve was achieved by varying, in concert the thresholds on all three classes of discrimination tree from 0.0 to 0.9. Figure 5 shows the trade-off curve achieved in EJV. The difference between the two curves highlights difference between the two domains and between the system configurations used for the two domains. The EME curve shows a much more dramatic trade-off. The EJV curve shows that only modest varying of recall and precision is achievable. Part of this is a reflection of the two domains. In EJV, most relationships were found via two noun phrases that shared a common CN trigger. This method proved to be effective at detecting relationships. Therefore the only real difference in the trade-off comes from varying the thresholds for the string-fill and set-fill trees, which generate the objects that are then composed into relationships. In EME, there not nearly as many shared triggers and so the template generator must attempt intelligent guesses for relations. The probabilistic guesses made in EME are much more amenable to threshold manipulation than the more structured information used in EJV. Also, in EJV the system ran with a slot masseur that embodied some domain knowledge. In EJV, TTG was configured to only hypothesize objects if the slot masseur had found a reasonable slot-fill or set-fill. This use of domain knowledge further limited the efficacy of changing certainty thresholds.

TRAINABLE INFORMATION EXTRACTION IN ACTION

Before CIRCUS can tackle an input sentence, we have to pass the source text through a preprocessor that locates sentence boundaries and reworks the source text into a list structure. The preprocessor replaces punctuation marks with special symbols and applies text processing specialists to pick up dates, locations, and other objects of interest to the target domain. We use the same preprocessing specialists for both EJV and EME: many specialists will apply to multiple domains. A subset of the Gazetteer was used to support the location specialist, but no other MRDs are used by the preprocessing specialists. We do not have a specialist that attempts to recognize company names.

Source Text:

```
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.
```

Preprocessed Text:

```
(*START* BRIDGESTONE SPORTS CO= SAID **ON_241189 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
*END*)
```

Note that the date specialist had to consult the dateline of the source text in order to determine that "Friday" must refer to November 24, 1989. Once the preprocessor has completed its analysis, the OTB part-of-speech tagger identifies parts of speech:

```
(*START* BRIDGESTONE SPORTS  CO=  SAID  **ON_241189 IT   HAS  SET  UP   A   JOINT
 strt    nm          nm      noun verb  $date$      noun aux  pasp ptcl art nm

VENTURE  IN   @@_Taiwan  WITH  A    LOCAL  CONCERN  AND  A    JAPANESE  TRADING
noun     prep $location$ prep  art  nm     noun     conj art  nm        nm

HOUSE  TO   PRODUCE  GOLF  CLUBS  TO   BE   SHIPPED  TO     @@_Japan    *END*)
noun   inf  verb     nm    noun   inf  aux  pasp     prep   $location$  stop
```

OTB tagged 97.1% of the words in EJV 0592 correctly. One error associated with "... A COMPANY ACTIVE IN TRADING WITH TAIWAN ..." led to a truncated noun phrase when "active" was tagged as a head noun instead of a nominative predicate.

With part-of-speech tags in place, CIRCUS can begin selective concept extraction. On this first sentence from EJV 0592, CIRCUS triggers 18 CN definitions triggered by the words "said" (3 CNs), "set" (3 CNs), "venture" (9 CNs), "produce" (1 CN), and "shipped" (2 CNs). These CNs extract a number of key noun phrases, and assign semantic features to these noun phrases based on soft constraints in the CN definition. Some of these features were recognized to be inconsistent with the slot fill and others were deemed acceptable. Notice that different CNs picked up "BRIDGESTONE SPORTS CO." with incompatible semantic features (it was associated with both a *joint venture* and a *joint venture parent* feature).

extracted NPs	rejected CN features:	accepted CN features:
BRIDGESTONE SPORTS CO=	jv-person	jv-entity, jv-parent, jv
IT	jv-person	
A JOINT VENTURE		jv-entity, company
GOLF CLUBS		jv-prod_serv, production

As we can see from this sentence, CN feature types are not always reliable, and CIRCUS does not always recognize the violation of a soft feature constraint. An independent set of semantic features are obtained from MayTag. In the first sentence of EJV 0592, MayTag only missed marking "golf clubs" as a product/service. An independent set of semantic features are obtained from MayTag:

words	MayTag semantic features	hits & misses
START		
BRIDGESTONE	((WS-JV-ENTITY) (WS-COMPANY-NAME))	; correct
SPORTS	((WS-JV-ENTITY) (WS-COMPANY-NAME))	; correct
CO=	((WS-JV-ENTITY) (WS-GENERIC-COMPANY))	; correct
SAID		
IT	((WS-ENTITY) NIL)	; correct
HAS		
SET		
UP		
A		
JOINT	((WS-JV-ENTITY) NIL)	; correct
VENTURE	((WS-JV-ENTITY) NIL)	; correct
IN		
JV-LOCATION	((WS-LOCATION) NIL)	; correct
WITH		
A		
LOCAL	((WS-ENTITY) NIL)	; correct
CONCERN	((WS-JV-ENTITY) NIL)	; correct
AND		
A		
JAPANESE	((WS-NATIONALITY) NIL)	; correct
TRADING	((WS-PRODUCT-SERVICE) (WS-SALES))	; correct
HOUSE	((WS-JV-ENTITY) NIL)	; correct
TO		
PRODUCE		
GOLF	((WS-ENTITY) NIL)	; incorrect
CLUBS	((WS-ENTITY) NIL)	; incorrect
TO		
BE		
SHIPPED		
TO		
JV-LOCATION	((WS-LOCATION) NIL)	; correct

In addition to extracting some noun phrases and assigning semantic features to those noun phrases, we also call the noun phrase classifier to see if any of the simple NPs picked up by the CN definitions should be extended to longer NPs. For this sentence, the noun phrase classifier extended only one NP: it decided that "A JOINT VENTURE" should be extended to pick up "A JOINT VENTURE IN TAIWAN WITH A LOCAL CONCERN AND A JAPANESE TRADING HOUSE". The second prepositional phrase should not have been included - this is an NP expansion that was overextended.

Each noun phrase extracted by a CIRCUS concept node will eventually be preserved in a memory token that records the CN features, MayTag features, any NP extensions, and other information associated with CN definitions. But before we look at the memory tokens, let's briefly review the other NPs that are extracted from the remainder of the text. For each preprocessed sentence produced in response to EJV 0592, we will put the noun phrases extracted by CIRCUS into boldface and use underlines to indicate how the noun phrase classifier extends some of these NPs.

START BRIDGESTONE SPORTS CO= SAID **ON_241189 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 *END*) (*START* THE JOINT VENTURE $COMMA$ BRIDGESTONE SPORTS TAIWAN CO= $COMMA$ CAPITALIZED AT $$20000000_TWD $COMMA$ WILL START PRODUCTION **DURING_0190 WITH PRODUCTION OF &&20000 IRON AND METAL WOOD CLUBS A MONTH *END*) (*START* THE MONTHLY OUTPUT WILL BE LATER RAISED TO &&50000 UNITS $COMMA$ BRIDGESTON SPORTS OFFICIALS SAID *END*) (*START* THE NEW COMPANY $COMMA$ BASED IN KAOHSIUNG $COMMA$ SOUTHERN TAIWAN $COMMA$ IS OWNED %%75 BY BRIDGESTONE SPORTS $COMMA$ %%15 BY UNION PRECISION CASTING CO= OF @@_Taiwan AND THE REMAINDER BY TAGA CO= $COMMA$ A COMPANY ACTIVE IN TRADING WITH TAIWAN $COMMA$ THE OFFICIALS SAID *END*) (*START* BRIDGESTONE SPORTS HAS SO_FAR BEEN ENTRUSTING PRODUCTION OF GOLF CLUB PARTS WITH UNION PRECISION CASTING AND OTHER TAIWAN COMPANIES *END*) (*START* WITH THE ESTABLISHMENT OF THE TAIWAN UNIT $COMMA$ THE JAPANESE SPORTS GOODS MAKER PLANS TO INCREASE PRODUCTION OF LUXURY CLUBS IN @@_Japan *END*)

As far as our CN dictionary coverage is concerned, we were able to identify all of the relevant noun phrases needed with the exception of "A LOCAL CONCERN AND A JAPANESE TRADING HOUSE" which should have been picked up by a JV parent CN. In fact, our AutoSlog dictionary had two such definitions in place for exactly this type of construction, but neither definition was able to complete its instantiation because of a previously unknown problem with time stamps inside CIRCUS. This was a processing failure - not a dictionary failure.

Trainable noun phrase analysis processes 13 of the 17 NP instances marked above correctly. Three of the NPs were expanded too far, and one was expanded but not quite far enough due to a tagging error by OTB ("a company active ..."). An inspection of the 13 correct instances reveals that 7 of these would have been correctly terminated by simple heuristics based on part-of-speech tags. It is important to note that the trainable NP analyzer had to deduce these more "obvious" heuristics in the same way that it deduces decisions for more complicated decisions. It is encouraging to see that straightforward heuristics can be acquired automatically by trainable classifiers. When our analyzer makes a mistake, it generally happens with the more complicated noun phrases (which is where hand-coded heuristics tend to break down as well).

After the noun phrase classifier has attempted to find the best termination points for the relevant NPs, we then call the coreference classifier to consider pairs of adjacent NPs separated by a comma. In this text we find three such appositive candidates (the second of which contains an extended NP that was not properly terminated):

```
THE JOINT VENTURE, BRIDGESTONE SPORTS TAIWAN CO.
TAGA CO., A COMPANY ACTIVE
THE NEW COMPANY, BASED IN KAOHSIUNG, SOUTHERN TAIWAN
```

In the third case, the location specialist failed to recognize either Kaohsiung or Southern Taiwan as names of locations. On the other hand, the fragment "based in Kaohsiung" was recognized as a location description and therefore reformatted it as "THE NEW COMPANY (%BASED-IN% KAOHSIUNG), SOUTHERN TAIWAN" which set up the entire construct as an appositive candidate. The coreference classifier then went on to accept each of these three instances as valid appositive constructions. This was the right decision in the first two cases, but wrong in the third. If full location recognition had been working, this last instance would have never been handed to the coreference classifier in the first place.

The coreference classifier tells us when adjacent noun phrases should be merged into a single memory token. We also invoke some hand-coded heuristics for coreference decisions that can be handled on the basis of lexical features alone. These heuristics determine that Bridgestone Sports Co. is coreferent with Bridgestone Sports, and that "THE JOINT VENTURE,, BRIDGESTONE SPORTS TAIWAN CO." is coreferent with "A JOINT VENTURE IN TAIWAN ..." Our lexical coreference heuristics are nevertheless very conservative, so they fail to merge our four product service instances in spite of the fact that "clubs" appears in three of these string fills. In effect, we pass the following memory token output to TTG:

5 recognized companies (#4 and #5 should have been merged):

```
(1) "TAGA CO=" aka "A COMPANY ACTIVE"
(2) "UNION PRECISION CASTING CO= OF @@_TAIWAN"
(3) "BRIDGESTONE SPORTS CO=" aka "BRIDGESTONE SPORTS"
(4) "THE NEW COMPANY (%BASED-IN% KAOHSIUNG)" aka "SOUTHERN TAIWAN"
(5) "THE JOINT VENTURE" aka "BRIDGESTONE SPORTS TAIWAN CO=" aka
    "A JOINT VENTURE IN @@_TAIWAN WITH A LOCAL CONCERN AND A JAPANESE TRADING HOUSE"
```

4 product service strings (all of these should have been merged):

```
(6) "GOLF CLUBS"
(7) "&&20000 IRON AND METAL WOOD CLUBS"
(8) "GOLF CLUB PARTS WITH UNION PRECISION CASTING AND OTHER TAIWAN COMPANIES"
(9) "LUXURY CLUBS IN @@_JAPAN"
```

1 ownership and 2 percent objects :

```
(10) "$$20000000_TWD"
(11) "%%15"
(12) "%%75"
```

We failed to extract "the remainder by ..." for the third ownership object because our percentage specialist was not watching for verbal referents in a percentage context - this could be fixed with an adjustment to the specialist.

When TTG receives memory tokens as input, the object existence classifiers try to filter out spurious information picked up by overzealous CN definitions. Unfortunately, in the case of 0592, TTG filtered out two good memory tokens: (#1 describing the parent Tago Co.), and (#5 describing the joint venture). It was particularly damaging to throw away #5 because that memory token contained the correct company name (Bridgestone Sports Taiwan Co.). Of the 3 remaining memory tokens describing companies, TTG correctly identified the two parent companies on the basis of semantic features, but then it was forced to pick up #4 as the child company. Our pathing function was smart enough to know that "THE NEW COMPANY" was probably not a good company name, but that left us with "SOUTHERN TAIWAN" for the company name. So a failure that started with location recognition led to a mistake in trainable appositive recognition, which then combined with a failure in lexical coreference recognition and a filtering error by TTG in order to give us a joint venture named "SOUTHERN TAIWAN" instead of "BRIDGESTON SPORTS TAIWAN." Overly aggressive filtering by TTG resulted in the loss of our 4 product service memory tokens.

Our CN instantiations do not explicitly represent relational information, but CNs that share a common trigger word can be counted on to link two CN instantiations in some kind of a relationship. Trigger families can reliably tell us when two entities are related, but they can't tell us what that relationship is. We relied on TTG to deduce specific relationships on the basis of its training. In cases like "75% BY BRIDGESTONE SPORTS", TTG had no trouble linking extracted percentage objects with companies. But our trainable link recognition ran into more difficulties when trigger families contained multiple companies Among the features that TTG had available for discrimination where closed class features, such as memory token types, semantic features, and CN patterns, and open class features (i.e. trigger words). However, although there exist heuristics for discriminating relationships based on particular words, the combination of the algorithms used (ID3) and the amount of data (600 stories) failed to induce these heuristics. The may be other algorithms, however, that used the same or less data and external knowledge to derive such heuristics from the training data.

The processing for EME proceeds very similarly to EJV, with the exception that MayTag is not used in our EME configuration, and in the EME system we used our standard CN mechanism and an additional keyword CN (KCN) mechanism. The KCN mechanism was used to recognize specific types of processing, equipment, and devices that have one or only a few possible manifestations. Below we see the OTB tags for the first sentence, all of which are correct. In fact, for EME text 2789568, OTB had 100% hit rate.

```
*start*  **DURING_2Q91      $COMMA$     Nikon  Corp=  $LPAREN$     &&7731
strt     $date$             punc        nm     noun   punc         noun

$RPAREN$  plans  to   market  the  NSR-1755EX8A  $COMMA$  a    new  stepper
punc      verb   inf  verb    art  noun          punc     art  nm   noun

intended  for   use   in    the  production  of    &&64-  Mbit  DRAMs  *end*
pasp      prep  noun  prep  art  noun        prep  nm     nm    noun   stop
```

The memory token structure below illustrates the processing of the text prior to TTG. Two NPs are identified as the same entity, "Nikon" and "Nikon Corp." The two NPs are merged into one memory token based on name merging heuristics. The second NP demonstrates how multiple recognition mechanisms can add robustness to the processing. "Nikon Corp." is picked up by both a CN triggered off of "plans to market" and by two KCNs, one that looks for "Corp." and another that looks for the lead NP in the story.

```
(TOKEN
 (TYPE (ME-ENTITY))
 (SUBTYPE NIL)
 (RELATION NIL)
 (SLOT-FILLS
   (TYPE COMPANY)
   (NAME (:SYM-LIST NIKON CORP=)))
 (NPS
   (NP 2 1 (NIKON)
     (CNS %ME-ENTITY-NAME-SUBJECT-VERB-AND-DO-STEPPER%))
   (NP 0 3 (NIKON CORP=)
     (CNS %ME-ENTITY-NAME-SUBJECT-VERB-AND-INFINITIVE-PLANS-TO-MARKET%)
     (KCNS %KEYWORD-ME-ENTITY-CORP=%
          %LEAD-NP%))))
```

Unfortunately, our system did not get any lithography objects for this story. On our list of things to get to if time permitted was creating a lithography object for an otherwise orphaned stepper. We would have only gotten one lithography object since we merged all mentions of "stepper" into one memory token.

We created a synthetic version of the system that inserted a lithography memory token corresponding to each stepper. One was discard by TTG and another was created because there were two different equipment objects attached to the remaining lithography object. The features that TTG used to hypothesize a new ME_CAPABILITY are illustrative of one of the weaknesses of this particular method. TTG used the following features to decide not to generate an ME_CAPABILITY developer.

FEATURE	RELATION CERTAINTY AFTER FEATURE
	0.36
The process is not X-RAY	0.23
The entity is not triggered off "developed"	0.14
The process is not CVD	0.03
The process is not LITHOGRAPHY of UKN type	0.04
The process is not ETCHING	0.06
The entity is not triggered off "from"	0.12

All of the features are negative, and the absence of each feature reduces the certainty that the relation holds, because each feature's presence, broadly speaking, is positive evidence of a relation. Therefore, the node of the decision tree that is found is a grouping of cases that have no particular positive evidence to support the relation, but also no negative evidence. With the relation threshold set at 0.3, this yields a negative identification of a relation. However, there are strong indications of a relation here. For example, the trigger "plans to market" is good evidence of a relation, however, the nature decision tree algorithms (recursively splitting the training data) causes us to loose that feature (in favor of other, better features). The following set of features shows what TTG used to generate an ME_CAPABILITY distributor.

FEATURE	RELATION CERTAINTY AFTER FEATURE
	0.38
The process is not packaging	0.47
The entity is not in a PP	0.58
A CN marked the entity as an entity	0.55
The process is not layering type sputtering	0.40

Again, we do not see here the features that we would expect, given the text. A human generating rules would say that "plans to market" is a good indication of a ME_CAPABILITY distributor.

DICTIONARY CONSTRUCTION BY DOMAIN EXPERTS

Sites participating in the recent message understanding conferences have increasingly focused their research on developing methods for automated knowledge acquisition and tools for human-assisted knowledge engineering. However, it is important to remember that the ultimate users of these tools will be domain experts, not natural language processing researchers. Domain experts have extensive knowledge about the task and the domain, but will have little or no background in linguistics or text processing. Tools that assume familiarity with computational linguistics will be of limited use in practical development scenarios.

To investigate practical dictionary construction, we conducted an experiment with government analysts. We wanted to demonstrate that domain experts with no background in text processing could successfully use the AutoSlog dictionary construction tool [Riloff and Lehnert 1993]. We compared the dictionaries constructed by the government analysts with a dictionary constructed by a UMass researcher. The results of the experiment suggest that domain experts can successfully use AutoSlog with only minimal training and achieve performance levels comparable to NLP researchers.

AutoSlog is a system that automatically constructs a dictionary for information extraction tasks. Given a training corpus, AutoSlog proposes domain-specific *concept node* definitions that CIRCUS [Lehnert 1991] uses to extract information from text. However, many of the definitions proposed by AutoSlog should not be retained in the permanent dictionary because they are useless or too risky. We therefore rely on a human-in-the-loop to manually skim the definitions proposed by AutoSlog and separate the good ones from the bad ones.

Two government analysts agreed to be the subjects of our experiment. Both analysts had generated templates for the joint ventures domain, so they were experts with the EJV domain and the template-filling task. Neither analyst had any background in linguistics or text processing and had no previous experience with our system. Before they began using the AutoSlog interface, we gave them a 1.5 hour tutorial to explain how AutoSlog works and how to use the interface. The tutorial included some examples to highlight important issues and general decision-making advice. Finally, we gave each analyst a set of 1575 concept node definitions to review. These included definitions to extract 8 types of information: jv-entities, facilities, person names, product/service descriptions, ownership percentages, total revenue amounts, revenue rate amounts, and ownership capitalization amounts.

We did not give the analysts all of the concept node definitions proposed by AutoSlog for the EJV domain. AutoSlog actually proposed 3167 concept node definitions, but the analysts were only available for two days and we did not expect them to be able to review 3167 definitions in this limited time frame. So we created an "abridged" version of the dictionary by eliminating jv-entity and product/service patterns that appeared only infrequently in the corpus.[3] The resulting "abridged" dictionary contained 1575 concept node definitions.

We compared the analysts' dictionaries with the dictionary generated by UMass for the final Tipster evaluation. However, the official UMass dictionary was based on the complete set of 3167 definitions originally proposed by AutoSlog as well as definitions that were spawned by AutoSlog's optional generalization modules. We did not use the generalization modules in this experiment, due to time constraints. To create a comparable UMass dictionary, we removed all of the "generalized" definitions from the UMass dictionary as well as the definitions that were not among the 1575 given to the analysts. The resulting UMass dictionary was a much smaller subset of the official UMass dictionary.

Analyst A took approximately 12.0 hours and Analyst B took approximately 10.6 hours to filter their respective dictionaries. Figure 6 shows the number of definitions that each analyst kept, separated by types. For comparison's sake, we also show the breakdown for the smaller UMass dictionary.

CN Type	# proposed by AutoSlog	# kept (UMass)	# kept (Analyst A)	# kept (Analyst B)
entity	688	311	357	423
facility	80	20	16	55
ownership-percent	174	91	117	91
person	243	119	149	52
prod_serv	316	76	152	44
revenue-rate	19	14	12	16
revenue-total	30	22	15	26
total-capitalization	25	14	13	22
TOTAL	1575	667	831	729

Figure 6: Comparative dictionary sizes

[3] While processing the training corpus, AutoSlog keeps track of the number of times that it proposes each definition (it may propose a definition more than once if the same pattern appears multiple times in the corpus). We removed all jv-entity definitions that were proposed < 2 times and all product/service definitions that were proposed < 3 times. We eliminated jv-entity and product/service definitions only because the sheer number of these definitions overwhelmed the other types.

We compared the dictionaries constructed by the analysts with the UMass dictionary in the following manner. We took the official UMass/Hughes system, removed the official UMass dictionary, and replaced it with a new dictionary (the smaller UMass dictionary or an analysts' dictionary). One complication is that the UMass/Hughes system includes two modules, TTG and Maytag, that use the concept node dictionary during training. In a clean experimental design, we should ideally retrain these components for each new dictionary. We did retrain the template generator (TTG), but we did not retrain Maytag. We expect that this should not have a significant impact on the relative performances of the dictionaries, but we are not certain of its exact impact. Finally, we scored each new version of the UMass/Hughes system on the Tips3 test set. Figure 7 shows the results for each dictionary.

TIPS3	Recall	Precision	P&R	ERR
UMass/Hughes	18	51	27.06	83
Analyst A	19	47	27.39	83
Analyst B	20	47	27.89	83

Figure 7: Comparative scores for Tips3

The F-measures (P&R) were extremely close across all 3 dictionaries. In fact, both analysts' dictionaries achieved slightly higher F-measures than the UMass dictionary. The error rates (ERR) for all three dictionaries were identical. But we do see some variation in the recall and precision scores. We also see variations when we score the three parts of Tips3 separately (see Figure 8).

TIPS3/Part1	Recall	Precision	P&R	ERR
UMass/Hughes	18	51	27.04	83
Analyst A	20	48	28.00	82
Analyst B	22	47	29.69	81

TIPS3/Part2	Recall	Precision	P&R	ERR
UMass/Hughes	17	52	26.03	84
Analyst A	18	48	25.92	84
Analyst B	20	47	27.75	83

TIPS3/Part3	Recall	Precision	P&R	ERR
UMass/Hughes	20	50	28.12	82
Analyst A	20	46	27.96	82
Analyst B	17	48	25.25	84

Figure 8: Comparative scores for Part1, Part2, and Part3

In general, the analysts' dictionaries achieved slightly higher recall but lower precision than the UMass dictionary. We hypothesize that this is because the UMass researcher was not very familiar with the corpus and was therefore somewhat conservative about keeping definitions. The analysts were much more familiar with the corpus and were probably more willing to keep definitions for patterns that they had seen before. There is usually a trade-off involved in making these decisions: a liberal strategy will often result in higher recall but lower precision whereas a conservative strategy may result in lower recall but higher precision.

It is interesting to note that even though there was great variation across the individual dictionaries (see Figure 6), the resulting scores were very similar. This may be because some definitions can contribute a disproportionate amount of performance if they are frequently triggered by a given test set. If the three dictionaries were in agreement on that subset of the dictionary that is most heavily used, those definitions could dominate overall system performance. Some dictionary definitions are more important than others.

To summarize, this experiment suggests that domain experts can successfully use AutoSlog to build domain-specific dictionaries for information extraction. With only 1.5. hours of training, two domain experts constructed dictionaries that achieved performance comparable to a dictionary constructed by a UMass researcher. Although this was only a small experiment, the results lend credibility to the claim that domain experts can build effective dictionaries for information extraction.

WHAT WORKS AND WHAT NEEDS WORK

When we look at individual texts and work up a walk through analysis of what is and is not working, we find that many of our trainable language components are working very well. The dictionary coverage provided by AutoSlog appears to be quite adequate. OTB is operating reliably enough for subsequent sentence analysis. When we run into difficulties with our trainable components, we often find that many of these difficulties stem from a mismatch of training data with test data. For example, when we trained the coreference interface for appositive recognition, we eliminated from the training data all candidate pairs involving locations because the location specialist should be identifying locations for us. If the coreference classifier were operating in an ideal environment, it would never encounter unrecognized locations. Unfortunately, as we saw with EJV 0592, the location specialist does not trap all the locations, and this led to a bad coreference decision. In an earlier version of the coreference classifier we had trained it on imperfect data containing unrecognized locations, but as the location specialist improved, we felt that the training for the coreference classifier was falling increasingly out of sync with the rest of the system so we updated it by eliminating all the location instances. Then when the coreference classifier was confronted with an unrecognized location, it failed to classify it correctly. When upstream system components are continually evolving (as they were during our MUC-5 development cycle), it is difficult to synchronize downstream dependencies in training data. A better system development cycle would stabilize upstream components before training downstream components in order to maintain the best possible synchronization across trainable components.

TTG was able to add some value to the output of CIRCUS and subsequent discourse processing. In module tests, TTG typically added 6-12% of accuracy in identifying domain objects and relationships. That added value is measured against picking that most likely class (yes or no) for a particular domain object (e.g. JV-ENTITY or ME-LITHOGRAPHY) or relationship (e.g. JV-TIE-UP or ME-MICROELECTRONICS-CAPABILITY). However, TTG fell far below our expectations for correctly filtering and connecting the parser's output. We find two reasons for this short fall. First, some small deficit can be attributed to the system development cycle since TTG sits at the end of the cycle of training and testing various modules.

The second, and by far the dominant effect comes from the combination of the training algorithm (ID3) and the amount of data. As mentioned previously, there are two types of features used by TTG: (1) closed class (e.g. token type, semantic features, and CN patterns) and (2) open class features (i.e. CN trigger words). Using open class features can be difficult, because most algorithms cannot detect reliable discriminating features if there are too many features—reliable features cannot be separated from noise. Using trigger words in conjunction relations between memory token results in 3,000-5,000 binary features. With no noise suppression added to the algorithm and given a large number of features, ID3 will create very deep decision trees that classify stories in the training set based on noise.

We ran two sets of decision trees in deciding how to configure our system for the final test run. MIN-TREES using only closed class features and no noise suppression and MAX-TREE using closed class and open class features and a noise suppression rule. The noise suppression was a termination condition on the recursion of the ID3 algorithm. Recursion was terminated when all features resulted in creating a node that classified examples from few than 10 different source texts. Using closed class features rarely resulted in a terminal node that classified examples from fewer than 10 stories. In all tests the MAX-trees performed better. However, as a result of the noise suppression, no decision tree contained very many discriminations on a trigger. The performance of the MAX-trees indicated that individual words are good discriminators, however their scarcity in the decision trees indicates that we are not using the appropriate algorithm. We believe that data-lean algorithms (such as explanation-based learning) in concert with shared knowledge bases might be effective.

In attributing performance to various components, we measured 25 random texts in EME. At the memory token stage we found that CIRCUS had extracted string-fills and set-fills with a recall/precision of 68/54. However our score output for those slots was 32/45 (measured only on the slots we attempted). Even when the thresholds for TTG were lowered to 0.0, so that all output came through, the recall was not anywhere near 68. Therefore it would appear that the difficult part of the template task is not finding good things to put in the template, but figuring how to split and merge objects. We do not (yet) have a trainable component that handles splitting and merging decisions in general.

The EJV and EME systems that we tested in our official evaluation were in many ways incomplete systems. Although our upstream components were operating reasonably well, additional feedback cycles were badly needed for other components operating downstream. In particular, trainable coreference and trainable template generation did not received the time and attention they deserve. We are generally encouraged by the success of our trainable components for part-of-speech tagging, dictionary generation, noun phrase analysis, semantic feature tagging, and coreference based on appositive recognition. But we encountered substantial difficulties with general coreference prior to template generation. This appears to be the greatest challenge remaining for trainable components supporting information extraction. We know from our earlier work in the domain of terrorism that coreference resolution can be reasonably well-managed on the basis of hand-coded heuristics [Lehnert et al. 1992b]. But this type of solution does not port across domains and therefore represents a significant system development bottleneck. True portability will only be achieved with trainable coreference capabilities.

We believe that trainable discourse analysis was the major stumbling block standing between our MUC-5 system and the performance levels attained by systems incorporating hand-coded discourse analysis. We remain optimistic that state-of-the-art performance will be obtained by corpus-driven machine learning techniques but it is clear that more research is needed to meet this very important challenge. To facilitate research in this area by other sites, UMass will make concept extraction training data (CIRCUS output) for the full EJV and EME corpora available to research laboratories with internet access. When paired with MUC-5 key templates available from the Linguistic Data Consortium, this data will allow a wide range of researchers who may not be experts in natural language to tackle the challenge of trainable coreference and template generation as problems in machine learning. We believe it is important for the NLP community to encourage and support the involvement of a wider research community in our quest for practical information extraction technologies.

BIBLIOGRAPHY

Cardie, C. (1993) A Case-Based Approach to Knowledge Acquisition for Domain-Specific Sentence Analysis. *Eleventh National Conference on Artificial Intelligence* (AAAI-93). Washington, D.C. pp. 798-803.

Dolan, C. P., Goldman, S. R., Cuda, T. V., & Nakamura, A. M. (1991). Hughes Trainable Text Skimmer: Description of the TTS System as Used for MUC-3. In B. Sundheim (Ed.), *Third Message Understanding Conference* (MUC-3). Naval Ocean Systems Center, San Diego California: Morgan Kaufmann. pp. 155-162.

Lehnert, W. (1991) Symbolic/Subsymbolic Sentence Analysis: Exploiting the Best of Two Worlds. *Advances in Connectionist and Neural Computation Theory. Vol. I.* (ed: J. Pollack and J. Barnden) Ablex Publishing, Norwood, New Jersey. pp. 135-164.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., Soderland, S. (1992a) University of Massachusetts: MUC-4 Text Results and Analysis *Proceedings of the Fourth Message Understanding Conference* (MUC-4). Morgan Kaufmann. San Mateo, CA. pp. 151-158.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., Soderland, S. (1992b) Description of the CIRCUS system as Used for MUC-4. *Proceedings of the Fourth Message Understanding Conference* (MUC-4). Morgan Kaufmann. San Mateo, CA. pp. 282-288.

Quinlan, J. R. (1983). Learning Efficient Classification Procedures and Their Application to Chess End Games. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Morgan Kaufmann. pp. 463-482.

Riloff, E. (1993) Automatically Constructing a Dictionary for Information Extraction Tasks. *Eleventh National Conference on Artificial Intelligence* (AAAI-93). Washington, D.C. pp. 811-816.

Riloff E., and Lehnert, W. (1993) Automated Dictionary Construction for Information Extraction from Text. *Proceedings of the Ninth IEEE Conference on Artificial Intelligence for Applications.* IEEE Computer Society Press. pp. 93-99.