Description of the LINK System Used for MUC-4

Steven L. Lytinen, Sayan Bhattacharyya, Robert R. Burridge,
Peter M. Hastings, Christian Huyck, Karen A. Lipinsky,
Eric S. McDaniel, and *Karenann K. Terrell*
Artificial Intelligence Laboratory
The University of Michigan
Ann Arbor, MI 48109
E-mail: lytinen@caen.engin.umich.edu

Background

The University of Michigan's natural language processing system, called LINK, is a unification-based system which we have developed over the last four years. Prior to MUC-4, LINK had been used to extract information from free-form texts in two narrow application domains. One application corpus contained terse descriptions of symptoms displayed by malfunctioning automobiles, and the repairs which fixed them. The other corpus described sequences of activities to be performed on an assembly line. In empirical testing in these two domains, LINK correctly processed 70% of previously unseen descriptions. A template was counted as correct only if all of the fillers in the template were filled correctly. In addition, LINK generated incomplete (but not incorrect) templates for another 15% of the descriptions.

These previous domains were much narrower than the MUC-4 terrorism domain. As a comparison, the lexicons for the previous domains contained only 300-500 words, compared with 6700 words in our MUC-4 test configuration. Previous grammar size ranged from 75-100 rules, compared with over 500 rules in the MUC-4 knowledge base. In addition, the previous application domains consisted only of single-sentence inputs. Thus, the integration of information from multiple sentences was not an issue in our previous work.

Flow of control

The MUC-4 LINK system consists of the modules shown in figure 1. One sentence at a time passes through the modules in the order shown in the figure. Each module's function is described below. To help explain the role of each module, its performance on various parts of message TST2-MUC4-0048 is shown.

The tokenizer

The tokenizer produces LISP-readable files from a 100-article source file. It also performs a few simple editing tasks, such as separating the text into sentences, and removing text that is in brackets.

The filter

The filter determines which sentences in an article should be passed to the remainder of the system for processing. While we originally had in mind more sophisticated filtering techniques, the filter in the test configuration simply passed on any sentences containing one or more words

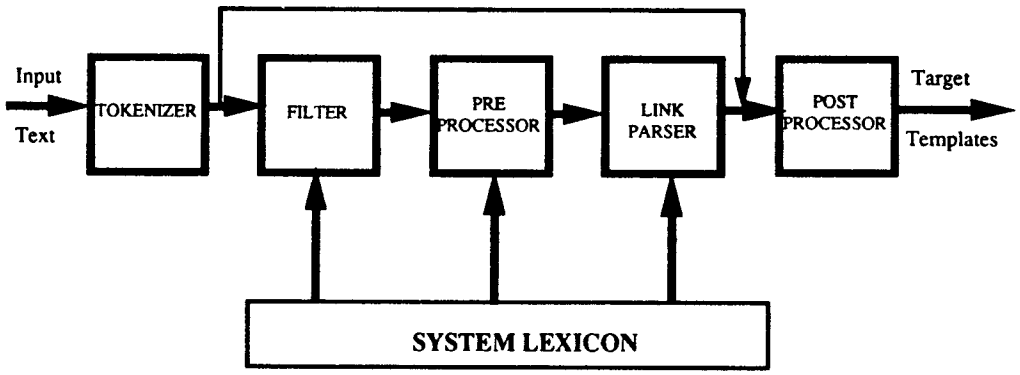

Figure 1: Modules of the MUC-4 LINK system

whose definitions were deemed interesting. Interesting definitions included any word meaning any of the template actions (BOMBING, ATTACK, ...) as well as a few other concepts likely to appear as template fillers, such as EXPLOSIVE and HOSTAGE. Any sentence which did not contain any words whose definitions were deemed interesting were discarded and not processed further.

The preprocessor

The preprocessor is responsible for initializing the environment in which the LINK parser operates. Since LINK is a bottom-up chart parser, this means that the preprocessor must initialize the chart. The initialization is constructed by looking up each word in the sentence, and adding a link into the chart corresponding to each possible sense of a word. In addition, likely noun phrases are identified and grouped together, and NP links are entered into the chart for these groups of words.

Noun phrases are preidentified for two reasons. First, undefined words often appear as parts of noun phrases. The grouping of these unidentified words eliminates the need to deal with them in the parser itself. Second, preprocessing noun phrases enabled us to encode parsing heuristics in the preprocessor which could not easily be encoded in the parser itself, such as preferring the longest possible noun phrase. This improves the efficiency of the system.

Each link in the initial chart contains both syntactic and semantic information about a word or noun phrase. For a single word, this semantic information is simply copied from the definition of the word. For noun phrases, semantic information on a link is the result of *unifying*, or merging together, semantic information from all of the defined words in the noun phrase. Adjacent nouns whose definitions cannot unify are not grouped together into a single noun phrase by the preprocessor. For example, "government headquarters" is not initially grouped as

a single NP, since the meanings of "government" and "headquarters" cannot be unified. Thus, it might be more accurate to say that the preprocessor identifies "noun clusters" rather than noun phrases.

If all of the words of an NP are undefined, then a default semantic definition is assigned. For the test configuration of the system, the default definition was HUMAN-OR-PLACE, a definition which could be refined during processing to be any of the set fills for the HUM TGT, PHYS TGT, or LOCATION fields.

The preprocessor is also responsible for identifying names of people. A list of names that appeared in the HUM TGT: NAME fields of the MUC-3 development answer keys is used to identify names, along with a few simple heuristics for identifying likely additional names. For example, any undefined word ending in a 'z' is considered a potential name.

Here are the results produced by the preprocessor for the first sentence in article TST2-MUC4-0048. The initial chart is displayed, with potential noun phrases already grouped together:

```
Next sentence (1):  SALVADORAN PRESIDENT-ELECT ALFREDO CRISTIANI CONDEMNED
    THE TERRORIST KILLING OF ATTORNEY GENERAL ROBERTO GARCIA ALVARADO AND
    ACCUSED THE FARABUNDO MARTI NATIONAL LIBERATION FRONT OF THE CRIME
Preprocessor results:

Node 0:  (SALVADORAN PRESIDENT-ELECT ALFREDO CRISTIANI)
Node 1:  CONDEMNED
Node 2:  (THE TERRORIST)
Node 3:  (KILLING) KILLING
Node 4:  OF
Node 5:  ((ATTORNEY GENERAL) ROBERTO GARCIA ALVARADO)
Node 6:  AND
Node 7:  ACCUSED
Node 8:  (THE (FARABUNDO MARTI NATIONAL LIBERATION FRONT))
Node 9:  OF
Node 10:  (THE CRIME)
Node 11:
```

The LINK parser

LINK is a bottom-up, unification-based chart parser. Its grammar rules are quite similar in form to those used in PATR-II (Shieber, 1986). We have incorporated semantic information into LINK's grammar, along the lines of HPSG (Pollard and Sag, 1987). The integration of syntactic and semantic knowledge into the same grammar formalism is crucial to our system's ability to process large texts in a reasonable length of time, and to producing the semantic analysis used to generate templates.

Here is a simplified example of a constraint rule:

```
(define-class S
      (1) = NP                                  <1>
      (2) = VP                                  <2>
      (head) = (2 head)                         <3>
      (head agr) = (1 head agr)                 <4>
      (head rep actor) = (1 head rep))          <5>
```

Each equation in this rule specifies a property which any node labeled S must have. A property consists of a *path*, or a sequence of arcs with the appropriate labels starting from the node in question; and a *value*, which is another node to be found at the end of the path. Equations specify the values of properties in one of two ways. They may specify the label of the node to be found at the end of the path, as in equations 1 and 2 (i.e., the arc from an S node labeled 1 leads to a node labeled NP). We will call these *labeling equations.* Or, they may specify that two paths must lead to the identical node, as in equations 3-5. Identity here is defined by the *unification* operation; i.e, if two paths must lead to the identical node, then the nodes at the end of the two paths must unify. Unification merges the properties of two nodes; thus, two paths can unify if their values have no properties which explicitly contradict each other. These equations will be called *unifying equations.*

Links are placed in the chart to represent potential constituents that the parser identifies. These links contain both syntactic and semantic information, represented in the form of a directed acyclic graph (DAG). The DAGs correspond to the information in the set of grammar rules used to build a constituent.

The core of the grammar is a set of domain-independent rules that handle all regular verb tenses, and many of the simple english constructions. The rules encode both syntactic and semantic constraints, which allows much of the work of finding the actor, object, location, etc. to be done during the parse.

The core is augmented by a set of rules that handle common constructions from the 1300 MUC-3 development articles. Typical examples of this are "Meanwhile, [sentence]." or "...an attack on [place]...", and are handled in as general a rule as possible provided the correct semantics may be given the parent based on the semantics of the children. An example of a grammar rule for a specific type of construction is shown below.

```
(define-class S
    ((1) = PRON
     (2) = VP2-PASS
     (3) = THAT
     (4) = S
     (1 1) = it
     (2 head rep) = strans
     (head) = (2 head)
     (head rep object) = (4 head rep)))
```

This rule handles all constructions of the form "It has been said that [sentence]" or "It was reported by the government today that [sentence]," etc.

Although the preprocessor is responsible for finding simple noun phrases, the correct interpretation of complex NPs relies on semantics and is handled by a set of grammar rules for NPs.

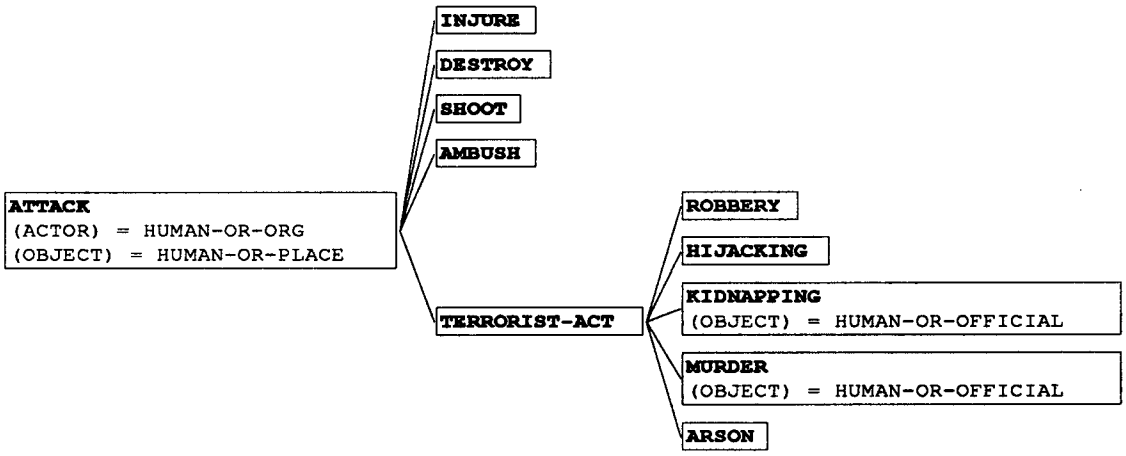

Figure 2: The ATTACK subtree of the concept hierarchy of actions for the terrorism domain

These include past particples used as adjectives (e.g., "the kidnapped priests"); noun phrase complements (e.g., "Noriega, the president of Panama") and some noun-noun constructions (e.g., "government headquarters" or "FMLN terrorists").

If a sentence fails to parse completely, the chart can be inspected to see what constituents have been constructed, and what their semantic content is. Thus, after a failed parse, the system examined the chart, identifying those links which contained information relevant to the construction of templates. Links which contained the most relevant information (i.e., the greatest number of slots filled which could map to template fields) were selected and passed to the postprocessor for incorporation into templates.

An example parse of sentence 1 from article TST2-MUC4-0048 is shown below.

```
((SENTENCE 1) "action1"
 ((ACTION-DESC (SEM-REF MURDER))
  (ACTOR (SEM-REF TERRORIST) (ACT-WORD (TERRORIST)))
  (OBJECT (SEM-REF GOVERNMENT-OFFICIAL) (ACT-WORD (ATTORNEY GENERAL))
   (NAME (ROBERTO GARCIA ALVARADO)))))
```

The Inheritance Hierarchy

The LINK parser utilizes semantic/domain knowledge during processing. This information is organized in an inheritance hierarchy. Figure 2 presents the actions from the hierarchy used in the MUC-4 domain, along with constraints on fillers of slots for actions. Slot-filling constraints on a concept may either be defined for that concept or inherited from the concept's ancestors in the tree. For example, since ATTACK requires an OBJECT that is a HUMAN-OR-PLACE, this restriction also implicitly holds for actions like SHOOT and ROBBERY. KIDNAPPING is an example of a concept which makes a further restriction on a previously constrained slot. HUMAN-OR-OFFICIAL, the OBJECT of this action, must be a descendant of HUMAN-OR-PLACE.

The Postprocessor

The postprocessor receives semantic representations from the LINK parser for each sentence in an article, and is responsible for producing response templates. It first checks to see if the representation of a sentence can be added to an existing template, or if it requires a new template. This decision is based on the compatibility of several template fields: the DATE, LOCATION, INCIDENT CATEGORY, and INDIVIDUAL ID fields. If all of these fields are compatible, then additional information is added to an existing template; otherwise, a new template is constructed.

To illustrate the strengths and weaknesses of the postprocessor, we will examine the processing of article TST2-MUC4-0048. Several sentences in this article generated templates. Here are the results produced by the LINK parser for sentences 1 and 2:

```
Sentence 1:  SALVADORAN PRESIDENT-ELECT ALFREDO CRISTIANI CONDEMNED THE
  TERRORIST KILLING OF ATTORNEY GENERAL ROBERTO GARCIA ALVARADO AND ACCUSED
  THE FARABUNDO MARTI NATIONAL LIBERATION FRONT OF THE CRIME

((SENTENCE 1) action1
 ((ACTION-DESC (SEM-REF MURDER))
  (ACTOR (SEM-REF TERRORIST) (ACT-WORD (TERRORIST)))
  (OBJECT (SEM-REF GOVERNMENT-OFFICIAL) (ACT-WORD (ATTORNEY GENERAL))
   (NAME (ROBERTO GARCIA ALVARADO)))))

Sentence 2:  LEGISLATIVE ASSEMBLY PRESIDENT RICARDO VALDIVIESO AND VICE
  PRESIDENT-ELECT FRANCISCO MERINO ALSO DECLARED THAT THE DEATH OF THE
  ATTORNEY GENERAL WAS CAUSED BY WHAT VALDIVIESO TERMED THE GUERRILLAS'
  IRRATIONAL VIOLENCE

((SENTENCE 2) action1
 ((ACTION-DESC (SEM-REF DIE))
  (RESULT action2 ((ACTION-DESC (SEM-REF NIL))))
  (LOCATION (SEM-REF NIL))
  (OBJECT (SEM-REF GOVERNMENT-OFFICIAL) (ACT-WORD (ATTORNEY GENERAL)))))
```

The templates for these sentences are merged, since the OBJECT of both actions appear to be the same person.

Later in the article, the following sentence appears:

```
Sentence 11: GUERRILLAS ATTACKED MERINO'S HOME IN SAN SALVADOR 5 DAYS AGO
  WITH EXPLOSIVES
```

```
((SENTENCE 11) action1
 ((ACTION-DESC (SEM-REF ATTACK))
  (ACTOR (SEM-REF TERRORIST) (ACT-WORD (GUERRILLAS)) (NUM PLURAL))
  (INSTRUMENT (SEM-REF EXPLOSIVE) (ACT-WORD (EXPLOSIVES)))
  (TIME (NUM 14) (MONTH APR) (SEM-REF DATE))
  (OBJECT (SEM-REF CIVILIAN-RESIDENCE) (ACT-WORD (MERINO'S HOME IN SAN SALVADOR))
   (LOCATION (SEM-REF CITY) (ACT-WORD (SAN SALVADOR))
      (COUNTRY EL-SALVADOR) (NAME (SAN-SALVADOR))))))
```

This information is not merged with the template generated from sentence 1 because of the mismatch between the OBJECTs of the two representations.

Sentence 22 illustrates the inability of our system to perform reference resolution:

```
Sentence 22:  ONE OF THEM WAS INJURED

((SENTENCE 22) action1 ((ACTION-DESC (SEM-REF DIE))))
```

Because the referent of 'them' cannot be resolved, it is dropped from the representation of the sentence, and the result is that no information is added to the response templates for this article.

References

Pollard, C., and Sag., I. (1987). *Information-based Syntax and Semantics.* Menlo Park, CA: Center for the Study of Language and Information.

Shieber, S. (1986). *An Introduction to Unification-based Approaches to Grammar.* CSLI, Stanford CA.