Scruffy Text Understanding: Design and Implementation of the NOMAD System

Richard H. Granger, Chris J. Staros,
Gregory B. Taylor, Rika Yoshii

Artificial Intelligence Project
Department of Information and Computer Science
University of California
Irvine, California 92717

ABSTRACT

The task of understanding unedited naval ship-to-shore messages is implemented in the presence of a large database of domain specific knowledge. The program uses internal syntactic and semantic expectations to analyze the texts and to correct errors that arise during understanding, such as syntactic errors, missing punctuation, and errors of spelling and usage. The output of the system is a well-formed English translation of the message. This paper describes some of the knowledge mechanisms that have been implemented in the NOMAD system.

I. Introduction

Consider the following message,

LOCKED ON OPEN FIRED DIW.

This is an actual naval message containing sentence boundary problems, missing subjects and objects, an incorrect verb conjugation, and an abbreviation for "dead in water." The NAVY receives many thousands of short messages like the one above in very 'scruffy' form, and these messages have to be put into a more readable form before they can be passed through many hands. Hence there is an obvious benefit to partially automating this encoding process.

Most large text-understanding systems today would not be able to automate the encoding process mentioned above because they were designed under the assumption that the input text consists of well-formed and logical sentences such as newspaper stories and other edited texts. The NOMAD system, however, was designed to understand naval text that contains ungrammatical or only partially complete sentences.

This paper explains some knowledge mechanisms that underlie the reader's ability to understand scruffy text and how these mechanisms are implemented within the NOMAD system.

This research was supported in part by the Naval Ocean Systems Center grant N00123-81-C-1078.

II. Categories of Errors

We have encountered the following problems in understanding Navy messages. They are listed in the order of frequency of their occurrences. The order was determined by examining the list of messages provided by the Naval Ocean Systems Center. For each error type, NOMAD's method of recognizing and correcting the problem are described, and the module which is responsible for the correction is identified.

A. Unknown words

Consider the message,

PEGASUS FRD 2 TALOS AT VICTOR

NOMAD does not immediately recognize 'FRD' as a word in the dictionary. Often the message sender will use an ad hoc abbreviation of a word or misspell a word. Any word not found to be in the dictionary is first put through a simple spelling correction procedure. If none of the possible corrections are recognizable then a morphological analyzer is applied to recognize different possible conjugations of a a verb.

If this fails, a mechanism called FOUL-UP (Granger, 1977) is triggered. The FOUL-UP mechanism handles unknown words by using the program's own syntactic and semantic expectations to create a temporary definition that would allow it continue normally. FOUL-UP would later revise the definition of the unknown word by combining the expectations generated based on previous information with the role the unknown word is playing in the current context.

B. Missing subject and objects

Consider the following message of two sentences,

CONSTELLATION SAW KASHIN. LOST CONTACT.

A script-based (Schank and Abelson, 1977) inferencer generates expectations to fill the subject and object of each sentence. Here, the word 'SAW' as a conjugation of 'SEE' would give arise to expectations related to detection and

identification. The inferencer also uses knowledge about typical sequences of events (identify before fire) (Cullingford, 1977) and relationships between their participants (friend and foe).

C. Ambiguous word usage

Examine the following message,

CONTACT GAINED ON KASHIN.

The example can be interpreted as either "Contact was gained on Kashin" meaning "We contacted Kashin" or "Our contact (a ship) made heading towards Kashin." NOMAD picks one of the multiple meanings of the ambiguous word, and calls a blame assignment module to check for goal violations, physical impossibilities, and other semantic conflicts to make sure that the interpretation was correct. If the module detects any conflict, NOMAD attempts to understand the sentence using a a different meaning of the ambiguous word.

D. Missing sentence and clause boundaries

Consider the following message,

VISUALLY LOCKED ON AND TRACKING CHALLENGED UNIT NO REPLY OPEN FIRED TIME 0129.1

NOMAD uses semantic expectations and syntactic expectations to detect missing boundaries. 'VISUALLY LOCKED ON' is understood to be a complete sentence because there are no expecations pending when 'AND' is read. 'TRACKing' is understood to be the verb of the second sentence. With a verb chosen and expecations for an actor pending, 'CHALLENGED' is used as an adverb describing 'UNIT'. The second phrase ends before 'NO REPLY ...' as again there are no expecations pending at this point. The phrase "NO REPLY" has expectations for communication verbs to follow it, and thus when the clause "OPEN FIRED" is encountered, the final sentence boundry is identified.

E. Wrong tense

Consider the following fragment sentence from our first example,

OPEN FIRED.

The morphological analyzer is used also to correct the tense of a word. eg. OPEN FIRED --> OPEN FIRE. The script-based inferencer then determines the tense of the given action using its knowledge about typical sequences of events. eg. LOCKED ON. OPEN FIRED. ---> LOCKED ON. OPENED FIRE.

III. Human Interface

NOMAD uses a generator specifically designed for the naval domain to produce a well formed translation of the input message. This 'pretty' form of the input message is checked by a user to assure that NOMAD has correctly understood the message. If NOMAD is then told it has incorrectly understood the message, alternative word definitions and other semantic choices are made in a second attempt at understanding.

A. The Generator

The generator has been tailored to address some of the problems that occur in describing naval activities [Taylor, 1982]. Many of the messages are characterized by incomplete and changing descriptions of actors. These descriptions contain information that may be unknown but doesn't change (name, type of ship, etc.) along with temporal information (location and damage status). The NOMAD generator produces unambiguous descriptions of actors while maintaining brevity.

IV. An Example

The following is an actual example showing the capability of NOMAD to handle multiple problems found in a message.

INPUT MESSAGE:

PERISCOPE SIGHTED BY CONSTELLATION ABT 2000 YDS OFF PORT QTR, AND HS HELO VECTRED TO DATUM. GREEN FLARES SIGHTED AFTER VISUAL ON PERISCOPE. HS GAINED ACTIVE CTC AND CONDUCTED TWO ATTACKS.

Two possible paraphrases are generated:

PARAPHRASE1:

The Constellation identified an enemy submarine that was at 225 degrees 2000 miles from their location. A helicopter-squadron pursued the enemy submarine. The helicopter-squadron identified some green flares. By using an active sonar, the helicopter-squadron identified the enemy submarine, and they fired twice at the enemy submarine.

PARAPHRASE2:

The Constellation identified an enemy submarine that was at 225 degrees 2000 miles from their location. A helicopter-squadron pursued the enemy submarine. The Constellation identified some green flares. By using an active sonar, the helicopter-squadron identified the enemy submarine, and they fired twice at the enemy submarine.

The main difference that is shown in the paraphrases is the identity of the subject of the second sentence. NOMAD gives preference in this case to the second paraphrase because 'AFTER VISUAL ON PERISCOPE' implies that the subject of the second sentence is the same as in the first sentence. However, the user is given the final choice.

V. Conclusions

The ability to understand text is dependent on the ability to understand what is being described in the text. Hence, a reader of, say, English text must have applicable knowledge of both the situations that may be described in texts (e.g., actions, states, sequences of events, goals, methods of achieving goals, etc.) and the the surface structures that appear in the language, i.e., the relations between the surface order of appearance of words and phrases, and their corresponding meaning structures.

The process of text understanding is the combined application of these knowledge sources as a reader proceeds through a text. This fact becomes clearest when we investigate the understanding of texts that present particular problems to a reader. Human understanding is inherently tolerant; people are naturally able to ignore many types of errors, omissions, poor constructions, etc., and get straight to the meaning of the text.

Our theories have tried to take this ability into account by including knowledge and mechanisms of error noticing and correcting as implicit parts of our process models of language understanding. The NOMAD system is the latest in a line of 'tolerant' language understanders, beginning with FOUL-UP, all based on the use of knowledge of syntax, semantics and pragmatics at all stages of the understanding process to cope with errors.

VI. REFERENCES

Cullingford, R. 1977. Controlling Inference in Story Understanding. Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, Mass.

Granger, R. 1977. FOUL-UP: A program that figures out meanings of words from context. Proceedings of the Fifth IJCAIL, Cambridge, Mass.

Schank, R. and Abelson R. 1977 Scripts, Plans, Goals and Understanding. Lawrence Erlbaum Associates, Hillsdale, N.J.

Taylor, G. 1982. English Generation Using More Than Just CDs. Internal NOMAD Design Documenation, UCI, 1982.

Wilensky, R. 1978. Undertanding Goal-based Stories. Computer Science Technical Report 140, Yale University.