PLANS FOR A TASK-ORIENTED EVALUATION OF NATURAL LANGUAGE UNDERSTANDING SYSTEMS

Beth M. Sundheim
Naval Ocean Systems Center, Code 444
San Diego, CA 92152-5000

ABSTRACT

A plan is presented for evaluating natural language processing (NLP) systems that have focused on the issues of text understanding as exemplified in short texts from military messages. The plan includes definition of bodies of text to use as development and test data, namely the narrative lines from one type of naval message, and definition of a simulated database update task that requires NLP systems to fill a template with information found in the texts. Documentation related to the naval messages and examples of filled templates have been prepared to assist NLP system developers. It is anticipated that developers of a number of different NLP systems will participate in the evaluation and will meet afterwards to present and interpret the results and to critique the test design.

INTRODUCTION

This project undertakes to provide meaningful measures of progress in the field of natural language processing (NLP). In particular, it is intended to result in definition of a theory- and implementation-independent test of the text analysis capabilities of text understanding systems that analyze short (paragraph-length) texts taken from military messages. The test is task-oriented in order to facilitate assessment of the general state of the art and provide a meaningful basis for comparing notes across systems. This design would seem to have two major problems, however: the reduction of a system's capabilities to a simple quantification of right versus wrong answers, and the lack of desired focus on understanding capabilities versus application capabilities.

It is claimed, however, that if the task performance is recorded on development data as well as test data and is repeated on the test data after updates are made, additional insights can be gained into a sytem's robustness, breadth and depth of coverage, and potential for handling novel text. A measurement of utility can be gained as well, by measuring performance on the original task, versus performance using a version of the inputs in which punctuation and spelling errors, highly elliptical constructions and sublanguage constructions have been eliminated. These additional measurements open up the black box to some extent, providing information that far exceeds what would be obtainable from a single measurement of performance on the test data in a blind test.

Also, despite the fact that the NLP systems are treated as black boxes, the evaluation should provide significant insights into their understanding versus application capabilities, because successful performance of the task does not require that back end modules contribute substantive information to the template fills. For example, the template fills do not require that any computations be performed on the data. This aspect of the test design is another way in which the black box has been opened up or narrowed down to increase the meaningfulness of the results. It is important, however, to recognize that the test is applicable only to "complete" and non-interactive systems, ones that are capable of accepting unseen texts and working essentially without human intervention to understand them.

This project is funded by DARPA/ISTO under ARPA Order No. 6359, Program Code 8E20, Program Element Code 62301E.

THE TEXT CORPUS

It is important that the amount of time and effort required for a system to be able to participate in the evaluation be as short as possible. For this reason, a serious effort has been made to collect texts in a narrow domain and to provide types of documentation that will reduce the amount of knowledge acquisition and engineering required. We have selected and prepared documentation on a set of 155 Navy messages written in a format known as OPREP-3 Pinnacle/Front Burner (OPREP-3 PFB), whose use and format are prescribed in OPNAVINST 3100.6D, "Special Incident Reporting," an unclassified Navy instruction. The examples selected concern encounters among aircraft, surface ships, and/or submarines. The encounters range in intensity from simple detection through overt hostility directed toward one of these "platform" types or an ashore facility. The nature of these messages is felt to be constrained in domain but not overly specialized.

OPREP-3 PFBs consist of several different paragraphs, each containing a prescribed type of information. The format of the information provided in each paragrph is generally unrestricted, and much of the information is supplied by message originators as free-form English text. The three major free-text paragraphs are (1) a narrative account of the incident, (2) a description of casualties suffered by personnel and equipment, and (3) miscellaneous remarks on the incident.

The OPREP-3 PFBs in the corpus have many features which make them tractable texts for current NLP systems:

1. They usually report on one or more closely-related general topics. The reported events fit into a fairly circumscribed set of scenarios concerning basic kinds of interaction between opposing forces of different types. Thus, the vocabulary is relatively limited, and so are the semantics of the domain.

2. They contain little speculation. At least in the narrative line, the author is attempting to report events as they occurred and not to speculate on those events. Thus, there is not too much in the way of complex constructions that convey an analysis (e.g. "[I] Believe that [the] attack was successful.").

3. They contain little embellishing information. They typically give only time, location and sensor/weapon information to supplement the recounting of the events. The succinct style preferred for Navy messages discourages the use of nonessential descriptive or qualifying expressions. This further reduces the number of different English constructions that a system would need to be able to syntactically parse, and restricts semantic interpretation mainly to representing fundamental attributes of agent, object, time, place, and instrument.

4. They stick basically to one topic per message. For the most part, it is not necessary to unravel a complex story, matching various events with different agents, objects, etc., and figuring out the time sequence.

Of course, there are also reasons why the text portions of OPREP-3 PFBs are in some ways very difficult to analyze. Some of the more superficial features that distinguish them from standard expository texts are

1. Poorer than average use of punctuation. Periods, especially, are sometimes omitted, leading to run-on sentences and increased amounts of ambiguity.

2. Heavy evidence of ellipsis (telegraphic style). Subjects, objects, articles, and prepositions are frequently omitted.

3. Use of special constructions, e.g., for representing time, date, and location.

4. Frequent misspellings. This is much more evident than in highly edited texts.

Some of the difficult distinguishing semantic features of OPREP-3 PFBs are

1. Assumption of knowledge of a specialized domain. The events, objects, and relationships in the Navy domain, e.g., what types of weapons can be used by what type of ship for what purpose, are not common knowledge. Frequently, the meaning of some part of a narrative will be somewhat ambiguous or vague to a nonspecialist, but completely clear to a knowledgeable person. Until a system developer has acquired a sound knowledge of the domain and has imparted it to the system's knowledge bases, the system is unlikely to perform any task very well.

2. Assumption of knowledge of contents of other paragraphs in the message. The narrative paragraphs are not intended to stand alone. The first paragraph of the message, for example, alerts the reader to the general subject of the message, so the narrative may omit some information that it would otherwise have included. That information may not be absolutely necessary for understanding the narrative in isolation but would help at least to reduce the degree of vagueness and ambiguity that the reader or system must resolve.

INPUTS TO NLP SYSTEMS: DEVELOPMENT AND TEST SETS

A total of 155 OPREP-3 PFBs are in the current corpus. Of these, 105 have been designated as development (i.e., training) data, and 50 have been set aside as test data. The current plan is to divide the test data into two sets of 25 messages each so that they can be used at different times in the future.

The corpus has been subdivided into four groups, according to the types of platforms involved in the interaction. There is one group each for incidents involving aircraft, surface ships, submarines, and land targets. The test data includes examples from each of these groups, in numbers proportional to the number of messages the development set contains for each group.

The inputs to the NLP systems are expected to be the OPREP-3 PFB narrative lines only. The intent is to limit the input to free text only and to about one paragraph in length. In that way, the task will focus on text understanding capabilities in general rather than on the understanding of a specialized message format, and it will include some, but not overwhelming, challenges for discourse-level processing.

As an alternative to the verbatim narrative lines, a set of modified versions is being prepared. The purpose is to allow systems that have not dealt extensively with the problems of telegraphic, often ill-formed texts to participate in the evaluation without having to undergo the extensive amount of development effort that would be required before they could be expected to have much success with the original narratives. Modifications will be made that minimize the superficial problems identified in the previous section (ellipsis, bad punctuation, specialized notation and misspellings). The evaluation of a system may be carried out using either the verbatim narratives or the modified versions, or both. For those systems which can analyze the unmodified inputs, a partial measurement of system utility can be obtained.

OUTPUTS FROM NLP SYSTEMS: DESCRIPTION OF THE TEMPLATE FILL TASK

The outputs are in the form of templates, simulating a simple database. No formal database management system is required. The software which must be developed especially for the benchmark test is a back end that takes the results of the analysis and extracts or derives the desired information to fill the slots in the template. This process is portrayed graphically below:

		NLP SYSTEM		NLP SYSTEM		
INPUTS		**FRONT END**		**BACK END**		**OUTPUTS**
OPREP-3 PFB NARRATIVES	**->**	**NL ANALYSIS MODULES**	**->**	**DATA EXTRACTOR/ DERIVER**	**->**	**TEMPLATE FILLS (DB UPDATES)**

The intention is that the back-end module required for the task be quite small and simple, since the test is meant to focus on the understanding capabilities, not on the sophistication of the system's database update capabilities. Systems will have to have mechanisms for mapping many kinds of data into canonical forms (see below), but there is no requirement for performing calculations on the data nor for other non-linguistic manipulation of the data.

The simulated database that will be created by the NLP systems is intended to capture basic information about events that are of significant interest. The events that will cause the system to fill in a template concern hostile or potentially hostile encounters between one or more members of the U.S. forces and one or more members of an enemy force -- detecting the enemy, tracking it, targeting it, harassing it, or attacking it. A template is also to be filled in if the action goes the opposite direction, i.e., where it is the enemy platform that is detecting, tracking, targeting, harassing, or attacking. Thus, the simulated database that is being created consists of the equivalent of two tables, one where the U.S. force carries out the action, and one where the enemy force carries out the action. Each time a new template is filled out, the equivalent of a new record is created for that table.

Not all OPREP-3 PFBs report one of the events mentioned above, however. There are some which report intentions rather than past events, and ones which report events that are "not of interest" to the database. Only the MESSAGE ID and EVENT slots (see below) should be filled out in these cases. This provides a check on the degree of understanding that a system is capable of, since there are times when a system that depended too heavily on key words, such as "attack," would mistakenly fill out a template.

SPECIFICATION OF THE TEMPLATE SLOTS

The template used in the benchmark test bears little resemblance to a comprehensive template schema such as that used by Logicon's Data Base Generator system for storing information on space event messages. It is intentionally simple, in an attempt to limit the amount of specialized back-end software the task requires, to limit the anticipated confusion and debate among system developers over what the expected "right answers" are, and to increase the comprehensibility of the output for all concerned. Unfortunately, by keeping the template simple, some specificity is lost that one would like to have in a database.

There are ten main slots in the template, plus one to identify the message that the data comes from. The slots and their fill requirements are given on the next page. The slots are meant to provide answers to the questions of What? Who? How? Where? When? With what outcome? The expected fill for each slot falls into one of two categories: selection of an item from a set list of possible answers, or strings (phrases) from the input text. As many of the fills as possible will come from predefined sets of possible names and categories. For the nomenclature identifying specific agents, objects, instruments, and locations, there will be correspondence tables that can be implemented to output a canonical form of identification.

Slot #1, which answers the question What?, is intended to indicate how serious the incident is by identifying the greatest level of hostility reported. In ascending order of hostility, the events are DETECT, TRACK, TARGET, HARASS, and ATTACK. The other possible fill for that slot is OTHER, meaning that the event is not of interest to the database. The remainder of the template should be left blank in that case. If the event is of interest to the database, the rest of the slots should be filled in; if information is not available for any of them, the phrase NO DATA should be given as the fill.

SLOT#	SLOT NAME	DATA FILL REQUIREMENTS
0	MESSAGE ID	From input header (DEV-GROUP1-N09722-001)
1	EVENT: HIGHEST LEVEL OF ACTION	DETECT, TRACK, TARGET, HARASS, ATTACK, OTHER
2	FORCE INITIATING EVENT	FRIENDLY, HOSTILE, NO DATA
3	CATEGORY(S) OF EVENT AGENT(S)	AIR, SURF, SUB, NO DATA
4	CATEGORY(S) OF EVENT OBJECT(S)	AIR, SURF, SUB, LAND, NO DATA
5	ID(S) OF 0-TH LEVEL AGENT(S)	Canonical form of name(s), else taxonomic category name(s) or organizational entity I.D., else NO DATA
6	ID(S) OF 0-TH LEVEL OBJECT(S)	Same as slot 5
7	INSTRUMENT(S) OF 0-TH AGENT(S)	Same as slot 5, where item(s) is/are: 1. sensor - for CONTACT, TRACK, TARGET 2. weapon - for HARASS, ATTACK
8	LOC OF OBJECT(S) AT EVENT TIME	Canonical form of location name(s), or text string with absolute or relative location(s), else NO DATA
9	TIME(S) OF EVENT	String with absolute time(s) of 1. use of sensor - for DETECT, TRACK, TARGET 2. weapon launch or impact - for HARASS, ATTACK; else NO DATA
10	RESULT(S) OF EVENT	1. RESPONSE BY OPPOSING FORCE 2. HOLDING CONTACT, LOST CONTACT 3. CONTINUING TO TRACK, STOPPED TRACKING 4. HOLDING TARGET, LOST TARGET 5. (NO) DAMAGE OR LOSS TO AGENT, (NO) DAMAGE OR LOSS TO OBJECT - 6. else, NO DATA

Table 1. Specifications for Output Template

A number of problems arose in preparing examples of filled templates, e.g., questions of how many templates were warranted and cases where the answers were unclear or did not fit the requirements exactly. On the other hand, there were many cases where the task showed promise of providing significant insights into the ability of NLP systems to correlate data, make inferences, filter out negative cases, and accommodate complex or ill-formed structures.

TEST PARAMETERS AND MEASURES

Several different measurements can be obtained from tests using the OPREP-3 PFB corpus. These can be termed "recall," "precision," "generality/potential," "utility," and "progress." The table below describes how measurements of them will be obtained and summarizes their significance as evaluation measures. Tests will be conducted by the system developers at their own sites at two different times. They will test the system upon receipt of 25 test narratives, which will come after a two-month period of updates for the development set. At that time, tests will be run separately for the development and test sets. After an additional month of updating to better handle the test set, the test will be rerun. As a final data point and stimulus for discussion, approximately 10 previously unseen narratives will be run by developers at the meeting following the period of updating. These narratives will be manufactured to be variations of narratives already seen, using the same situations and terminology in novel ways.

MEASURE	MEANS OF CALCULATION	INSIGHTS GAINED
RECALL	1) Percent templates for test set filled that should be filled (a template is to be filled in only if the message reports a past DETECT, TRACK, TARGET, HARASS or ATTACK event) 2) Percent slots for test set filled that should be filled	1) Coverage of English 2) Depth of understanding 3) Robustness
PRECISION	Of those slots correctly recalled for test set, percent correctly filled	Same as for RECALL
GENERALITY/ POTENTIAL	Comparison of recall and precision measures for development vs test set	1) Generality of software 2) Potential for handling novel text
UTILITY	Recall and precision measurements using original narratives as input (vs measurements obtained for modified narrative inputs)	Ability to handle telegraphic, ill-formed text and sublanguage constructions
PROGRESS	Results of run using first test set compared to results obtained later using second test set	Pace at which state of the art is advancing

Table 2. Evaluation Criteria and Significance

SUMMARY

This evaluation project represents a small step in the direction of benchmarking NLP systems. The results of the evaluation are not expected to be statistically significant but will begin to bring quantitative criteria to bear. They will also provide common ground for discussion of the language processing issues raised by military texts and the task-related issues of extracting, deriving and inferring needed information for a database. Developers of a number of different NLP systems that have processed military message texts are being invited to participate in this evaluation effort and will meet afterward to discuss the results and critique the test design. Although the test is designed to treat the NLP systems as black boxes, it is expected to yield a variety of meaningful measurements. It is also expected that discussion of the test results at the end of the evaluation will provide a great deal of insight not only into what systems can do but also how they do it.

REFERENCES

Department of the Navy. Special Incident Reporting. OPNAVINST 3100.6D.

Montgomery, C., and Glover, B. (1986). A Sublanguage for Reporting and Analysis of Space Events. In R. Grishman and R. Kittredge (Eds.), Analyzing Language in Restricted Domains: Sublanguage Description and Processing (pp. 129-161). Hillsdale, NJ: Erlbaum.

Sundheim, B. (1989). Navy Tactical Incident Reporting in a Highly Constrained Sublanguage: Examples and Analysis. Naval Ocean Systems Center Technical Document 1477 (in press).