Ill-Formed and Non-Standard Language Problems

Stan Kwasny
Computer Science Department
Indiana University
Bloomington, IN 47405

Abstract

Prospects look good for making real improvements in Natural Language Processing systems with regard to dealing with unconventional inputs in a practical way. Research which is expected to have an influence on this progress as well as some predictions about accomplishments in both the short and long term are discussed.

1. Introduction

Developing Natural Language Understanding systems which permit language in expected forms in anticipated environments having a well-defined semantics is in many ways a solved problem with today's technology. Unfortunately, few interesting situations in which Natural Language is useful live up to this description. Even a modicum of machine intelligence is not possible, we believe, without continuing the pursuit for more sophisticated models which deal with such problems and which degrade gracefully (see Hayes and Reddy, 1979).

Language as spoken (or typed) breaks the "rules". Every study substantiates this fact. Malhotra (1975) discovered this in his studies of live subjects in designing a system to support decision-making activities. An extensive investigation by Thompson (1980) provides further evidence that providing a grammar of "standard English" does not go far enough in meeting the prospective needs of the user. Studies by Fromkin and her co-workers (1980), likewise, provide new insights into the range of errors that can occur in the use of language in various situations. Studies of this sort are essential in identifying the nature of such non-standard usages.

But more than merely anticipating user inputs is required. Grammaticality is a continuum phenomenon with many dimensions. So is intelligibility. In hearing language used in a strange way, we often pass off the variation as dialectic, or we might unconsciously correct an errorful utterance. Occasionally, we might not understand or even misunderstand. What are the rules (meta-rules, etc.) under which we operate in doing this? Can introspection be trusted to provide the proper perspectives? The results of at least one investigator argue against the use of intuitions in discovering these rules (Spencer, 1973). Computational linguists must continue to conduct studies and consider the results of studies conducted by others.

2. Perspectives

Several perspectives exist which may give insights on the problem. We present some of these, not to pretend to exhaustively summarize them, but to hopefully stimulate interest among researchers to pursue one or more of these views of what is needed.

Certain telegraphic forms of language occur in situations where two or more speakers of different languages must communicate. A pidgin form of language develops which borrows features from each of the languages. Characteristically, it has limited vocabulary and lacks several grammatical devices (like number and gender, for example) and exhibits a reduced number of redundant features. This phenomenon can similarly be observed in some styles of man-machine dialogue. Once the user achieves some success in conversing with the machine, whether the conversation is being conducted in Natural Language or not, there is a tendency to continue to use those forms and words which were previously handled correctly. The result is a type of pidginization between the machine dialect and the user dialect which exhibits pidgin-like characteristics: limited vocabulary, limited use of some grammatical devices, etc. It is therefore reasonable to study these forms of language and to attempt to accomodate them in some natural way within our language models. Woods (1977) points out that the use of Natural Language:

> "... does not preclude the introduction of abbreviations and telegraphic shorthands for complex or high frequency concepts -- the ability of natural English to accommodate such abbreviations is one of its strengths." (p.18)

Specialized sublanguages can often be identified which enhance the quality of the communication and prove to be quite convenient especially to frequent users.

Conjunction is an extremely common and yet poorly understood phenomenon. The wide variety of ways in which sentence fragments may be joined argues against any approach which attempts to account for conjunction within the same set of rules used in processing other sentences. Also, constituents being joined are often fragments, rather than complete sentences, and, therefore, any serious attempt to address the problem of conjunction must necessarily investigate ellipsis as well. Since conjunction-handling involves ellipsis-handling, techniques which treat non-standard linguistic forms must explicate both.

3. Techniques

What approaches work well in such situtations? Once a non-standard language form has been identified, the rules of the language processing component could simply be expanded to accomodate that new form. But that approach has limitations and misses the general phenomenon in most cases.

DeJong (1979) demonstrated that wire service stories could be "skimmed" for prescribed concepts without much regard to grammaticality or acceptability issues. Instead, as long as coherency existed among the individual concepts, the overall content of the story could be summarized. The whole problem of addressing what to do with non-standard inputs was finessed because of the context.

Techniques based on meta-rules have been explored by various researchers. Kwasny (1980) investigated specialized techniques for dealing with cooccurrence violations, ellipsis, and conjunction within an ATN grammar. Sondheimer and Weischedel (1981) have generalized and refined this approach by making the meta-rules more explicit and by designing strategies which manipulate the rules of the grammar using meta-rules.

Other systems have taken the approach that the user should play a major role in exercising choices about the interpretations proposed by the system. With such feedback to the user, no time-consuming actions are performed without his approval. This approach works well in database retrieval tasks.

4. Near and Long Term Prospects

In the short term, we must look to what we understand and know about the language phenomena and apply those techniques that appear promising. Non-standard language forms appear as errors in the expected processing paths.

One of the functions of a style-checking program (for example the EPISTLE system by Miller et al., 1981) is to detect and, in some cases, correct certain types of errors made by the author of a document. Since such programs are expected to become more of a necessary part of any author support system, a great deal of research can be expected to be directed at that problem.

A great deal of research which deals with errors in language inputs comes from attempts to process continuous speech (see, for example, Bates, 1976). The techniques associate with non-left-to-right processing strategies should prove useful in narrowing the number of legal alternatives to be attempted when identifying and correcting some types of error. It is quite conceivable that an approach to this problem that parallels the work on speech understanding would be very fruitful. Note that this does not involve inventing new methods, but rather borrows from related studies. The primary impediment, at the moment, to this approach, as with some of the other approaches mentioned, is the time involved in considering viable alternatives. As these problems are reduced over the next few years, I feel that we should see Natural Language systems with greatly improved communication abilities.

In the long term, some form of language learning capability will be critical. Both rules and meta-rules will need to be modifiable. The system behavior will need to improve and adapt to the user over time. User models of style and preferred forms as well as common mistakes will be developed as a necessary part of such systems. As speed increases, more opportunity will be available for creative architectures such as was seen in the speech projects, but which still respond within a reasonable time frame.

Finally, formal studies of user responses will need to be conducted in an ongoing fashion to assure that the systems we build conform to user needs.

5. References

Bates, M., "Syntax in Automatic Speech Understanding," American Journal of Computational Linguistics, Microfiche 45, 1976.

DeJong, G.F., "Skimming Stories in Real Time: An Experiment in Integrated Understanding," Technical Report 158, Yale University, Computer Science Department, 1979.

Fromkin, V.A., ed., Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand, Academic Press, New York, 1980.

Hayes, P.J., and R. Reddy, "An Anatomy of Graceful Interaction in Spoken and Written Man-Machine Communication," Technical Report, Carnegie-Mellon University, August, 1979.

Kwasny, S.C., "Treatment of Ungrammatical and Extra-grammatical Phenomena in Natural Language Understanding Systems," Ph.D. Thesis, Ohio State University, 1980, (available through the Indiana University Linguistics Club, Bloomington, Indiana).

Kwasny, S.C., and N.K. Sondheimer, "Relaxation Techniques for Parsing Ill-Formed Input," American Journal of Computational Linguistics, Vol. 7, No. 2, April-June, 1981, 99-108.

Malhotra, A., "Design Criteria for a Knowledge-Based English Language System for Management: An Experimental Analysis," MAC TR 146, Cambridge, MA, M.I.T., February, 1975.

Miller, L.A., G.E. Heidorn, and K. Jensen, "Text-Critiquing with the EPISTLE System: An Author's Aid to Better Syntax," Proceedings of the National Computer Conference, AFIPS Press, Montvale, NJ, 1981.

Sondheimer, N.K., and R.M. Weischedel, "A Computational Linguistic Approach to Ungrammaticality Based on Meta-Rules" Annual Meeting of the Linguistic Society of America New York, NY, December, 1981.

Spencer, N.J., "Differences Between Linguists and Nonlinguists in Intuitions of Grammaticality-Acceptability" Journal of Psycholinguistic Research, 2, 2, 1973, 83-99.

Thompson, B.H., "Linguistic Analysis of Natural Language Communication with Computers," Proceedings of the Eighth International Conference on Computational Linguistics, Tokyo, October, 1980, 190-201.

Weischedel, R.M., and N.K. Sondheimer, "A Framework for Processing Ill-Formed Input," Technical Memorandum H-00519, Sperry-Univac, Blue Bell, PA, October 16, 1981.

Woods, W.A., "A Personal View of Natural Language Understanding," SIGART Newsletter, No. 61, February, 1977, 17-18.