HANDLING ILL-FORMED INPUT: SESSION INTRODUCTION

Ralph M. Weischedel
Department of Computer and Information Sciences
University of Delaware
Newark, Delaware 19711

1. Introduction

Suppose we call "normative" any system based on a set of constraints (whether pragmatic, semantic, or syntactic). Input that violates the constraints of a system could be termed "ill-formed". Sondheimer and Weischedel (1980) identify two general classes of input appearing ill-formed to a normative system. Input will be called

> absolutely ill-formed, if native speakers generally agree that it violates one or more linguistic constraints, or
> relatively ill-formed, if it violates some constraint(s) of the computational system, though native speakers perceive nothing odd about it.

Examples of absolute ill-formedness include misspellings, mistypings, mispunctuations, tense and number errors, word order problems, run-on sentences, sentence fragments, extraneous forms, and meaningless sentences. Examples of relatively ill-formed input include unknown words and requests that are beyond the limits of either the computer system or the natural language interface.

In natural language access (e.g. English access) to information systems, the magnitude of the problem of absolute ill-formedness can be seen in several case studies. If one includes telegraphic and elliptical constructions in the class of absolute ill-formedness, then case studies reported in Thompson (1980) and Eastman and McLean (1981) indicate that as much as 25% of queries to question-answering systems are absolutely ill-formed. On the other hand, no matter how large the dictionary, grammar, and underlying system, there will always be unknown words and phrases (e.g. proper names) and impossible requests (due to user misconceptions of the capabilities of the underlying system).

2. Brief overview of the papers of this session

This session consists of papers by Jensen and Heidorn; Marsh; and Granger et al. The paper by Jensen and Heidorn presents a particular heuristic for dealing with unparsable input. Since they have separate explicit heuristics for specific ungrammatical forms, a significant proportion of unparsable input in their system will be relatively ill-formed. One of the goals of the EPISTLE project, of which this is a part, is critiquing business letters.

Marsh's paper describes a technique for filling in material omitted from fragmentary inputs. Both syntactic information and domain-specific constraints on semantic classes are used. This is part of an ongoing effort in extracting a data base from free-text medical records, such as narrative discharge summaries.

The paper by Granger et al. reports on NOMAD, a system for taking cryptic, errorful naval ship-to-shore messages and generating well-formed versions. The paper describes the methods used for processing unknown words, fragments, missing punctuation, and tense errors.

One way to classify ill-formedness work is by the choices made regarding several issues. Each of the following sections will present one issue.

*This material is based upon work partially supported by the National Science Foundation under Grant No. IST-8009673.

3. Application area and responding to ill-formedness

Ill-formedness processing has been examined in many application areas, including data base access (Hendrix, et al., 1978), building a data base (Marsh, 1983), grammar checking (Jensen and Heidorn, 1983),. generating well-formed messages from ill-formed and incomplete ones (Granger, et al., 1983) and intelligent computer-assisted language instruction (Weischedel, et al., 1978).

Proper response to an ill-formed input depends on the application environment and the user. For example, suppose a presupposition of an input is incorrect according to the computer model. In a language instruction environment, the system should correct the erroneous input, informing the student of the cause of the error; this makes sense since students are error prone in learning a language. In data base access, the system might inform the user of the false presupposition and suggest alternative queries (Kaplan, 1978), since the user does not usually benefit from the empty set as a response. On the other hand, in the application Granger is investigating, the system should check an assumption it has made regarding the incomplete text and try an alternative, since the inconsistency may stem from an assumption the system made in completing fragmentary input.

4. The role of constraints

Perhaps ill-formedness should be handled by simply not encoding some constraints in the normative system. Waltz (1978) and Schank, et al. (1980) have taken this approach for a large class of syntactic constraints. In certain applications, one can get by with such an approach due to the great amount of redundancy in natural language. However, it clearly will not work in general, since constraints generally carry meaning and trim the search space. All three papers of this session capitalize on constraints (or expectations), and propose mechanisms for processing ill-formedness.

Given a commitment to employ constraints rather than ignoring them, there are still the following design issues:

a) whether two separate systems, perhaps running in parallel, should be built for well-formedness and for ill-formedness
b) whether ill-formed processing can be stated as explicit modifications to well-formedness processing, and
c) whether a metric can be used instead of employing any special procedures for ill-formedness.

Weischedel and Sondheimer (1981) argue for an approach that explicitly relates ill-formedness processing to the rules of the normative system via meta-rules (rules that operate on rules). A more complete discussion of the alternatives appears there.

5. Categorization of ellipsis, conjunction, and other gray cases

It is debatable whether certain phenomena should be classified as well-formed or not. Ellipsis (fragmentary input which in context conveys a complete thought) is an example. In such cases, the definition of "ill-formedness" in terms of a normative system, as well as the distinction between absolute and relative ill-formedness, sheds light on the issue. In Marsh's system, syntactic forms for ellipses are explicitly coded in the normative system. Jensen and Heidorn do not include rules for fragments in the grammar, but view it as relative ill-formedness to be processed by a recovery strategy. Granger, et al. also view ellipsis as relatively ill-formed.

Conjunction formation is another interesting case. Though use of conjunction is clearly grammatical, a number of individuals (Sager, 1973; Woods, 1973; Kwasny and Sondheimer, 1981) have argued that it should not be included in the normative grammar, but rather should be processed via a recovery strategy. Therefore, they have argued for treating it as relatively ill-formed.

I suspect that categorizing almost any particular constraint as normative could be the basis for argument. The criteria for deciding whether a constraint should be included in the normative system include at least the following:

a) whether a native speaker would edit inputs that violate it,
b) whether violating the constraint can yield useful inferences,
c) whether examples exist where the constraint carries meaning,
d) whether the constraint, if classified as normative, trims the search space, and
e) whether a processing strategy for the constraint can be stated more easily as a modification of normative processing, as in the case of conjunction.

6. Knowledge sources

In processing ill-formedness, there is usually more than one alternative for diagnosing the problem and for recovering. Oftentimes there is more than one alternative for what was intended. What kinds of knowledge are brought to bear on the problem? Jensen and Heidorn use syntactic information and an ordering heuristic in their parser. Marsh uses syntactic information and semantics (primarily selection restrictions). Granger et al. also employ syntactic and semantic constraints, but additionally employ "scripts" of stereotypical events in the environment of naval ship-to-shore messages.

Using more classes of information to infer what was intended is an open problem; the kinds of semantic and pragmatic knowledge that could be helpful have barely been tapped.

7. Control

The control mechanisms of the normative system (e.g. bottom-up versus top-down and the point at which semantic constraints are used) are not of concern to us here. Rather, what is of interest is

a) the point at which ill-formedness strategies are employed,
b) the mechanism for identifying what the problem is,
c) the nature of response, if any, once a specific hypothesis of the problem is made, and
d) the search strategy for selecting an intended interpretation.

Many alternatives exist for these decisions; an overview of them appears in Sondheimer and Weischedel (1980).

8. Future directions

As evident from this session, the processing of ill-formed input has become a very active research topic that is of critical importance in a wide variety of applications. Yet, there is much to be done. There are many kinds of ill-formedness for which better heuristics are needed.

Another need is empirical studies. Controlling the processing of ill-formed input is a substantial problem no matter what approach one takes, since processing ill-formedness requires relaxing the very constraints that trim the search space for possible interpretations. Because control is such an important issue, thorough, rigorous empirical studies are much needed to determine the effectiveness and costs of competing strategies.

Publishing additional collections of ill-formed input is critical. The patterns of behavior evident in such collections not only suggest heuristics for ill-formedness processing but also provide a basis for benchmarks upon which to base empirical comparisons.

One other area needing much research is models of particular users, their plans, and goals. This is important to infer the intended meaning of an individual, since many explanations exist. For instance, when no interpretation can be found, spelling corrections, inferring the meaning of an unknown word, or relaxing a syntactic or semantic constraint are all possibilities. Granger, et al. (1983) and Allen, et al. (1983) both report progress on using pragmatic information to deal with fragmentary input.

The obvious benefit of ill-formedness research is more natural, easy-to-use systems. An additional benefit is that study of ill-formedness should lead to better understanding of how normative systems should be designed.

References

Allen, James F., Alan M. Frisch, Diane J. Litman, "ARGOT: The Rochester Dialogue System", Proceedings of the National Conference on Artificial Intelligence, (1982), 66-70.

Eastman, C. M. and D. S. McLean, "On the Need for Parsing Ill-formed Input," American Journal of Computational Linguistics, 7, 4, (1981), 257.

Granger, Richard H., Chris J. Staros, Gregory B. Taylor, and Riku Yoshii, "Scruffy Text Understanding: Design and Implementation of the NOMAD System", this volume, 1983.

Hendrix, G. G., E. D. Sacerdoti, D. Sagalowicz and J. Slocum, "Developing a Natural Language Interface to Complex Data", ACM Transactions on Database Systems, 3, 2, (1978), 105-147.

Jensen, Karen and George E. Heinorn, "The 'Fitted' Parse: 100% Parsing Capability in a Syntactic Grammar of English", this volume, 1983.

Kaplan, S. J., "Indirect Responses to Loaded Questions," Theoretical Issues in Natural Language Processing-2, University of Illinois at Urbana-Champaign, July, 1978.

Kwasny, S. C., and N. K. Sondheimer, "Relaxation Techniques for Parsing Ill-Formed Input", American Journal of Computational Linguistics, 7, (1981), 99-108.

Marsh, Elaine, "Utilizing Domain-Specific Information for Processing Compact Text", this volume, 1983.

Sager, Naomi, "The String Parser for Scientific Literature." In R. Rustin, Ed., Natural Language Processing. New York: Algorithmics Press, 1973.

Schank, Roger C., Michael Lebowitz, and Lawrence Birnbaum, "An Integrated Understander", American Journal of Computational Linguistics, 6, 1, (1980), 13-30.

Sondheimer, N. K. and R. M. Weischedel, "A Rule-Based Approach to Ill-Formed Input," Proceedings of the Eighth International Conference on Computational Linguistics, Tokyo, October 1980, 46-54.

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.

Waltz, D. L., "An English Language Question Answering System for a Large Relational Database", Communications of the ACM, 21, 7, (1978), 526-539.

Weischedel, R.M. and N. K. Sondheimer, "A Framework for Processing Ill-Formed Input", Dept. of Computer & Information Sciences, University of Delaware, Newark, DE, 1981.

Weischedel, R. M., W. M. Voge, and M. James, "An Artificial Intelligence Approach to Language Instruction", Artificial Intelligence, 10, (1978), 225-240.

Woods, W. A., "An Experimental Parsing System for Transition Network Grammars." In R. Rustin, Ed., Natural Language Processing. New York: Algorithmics Press, 1973.