THE BOUNDARIES OF LANGUAGE GENERATION

Neil M. Goldman
USC - Information Sciences Institute

I. INTRODUCTION

In this paper I would like to address several basically independent issues relating to the processes of natural language generation (NLG) and research on modeling these processes. In the subsection "Paradigms for Generation" I maintain that, viewed at a moderately abstract level, the vast majority of current research in this area falls into a single model and focuses on the 'tail end' of the language generation process. The difference between individual models seems to be based on differing assumptions or convictions regarding the nature of 'pre-generative' aspects of language use.

The subsection 'Conceptual Generation' describes the particularized version of this basic model within which I work. The assumptions underlying this approach and the aspects of language generation which it attempts to account for are stressed.

The discussion of "Generation and Understanding" addresses the question of why a heavy bias can be seen in the volume of work (at least in the fields of computational linguistics and Artificial Intelligence) on language understanding as opposed to language generation. A related question - whether the two parts of language processing are sufficiently different to warrant independent stature - is discussed briefly.

The conclusion of the paper points up some of the areas of inquiry which have scarcely been touched, yet which must be developed before we can claim to have a model of the overall process of language generation.

II. PARADIGMS FOR GENERATION

A straightforward interpretation of the term "natural language generation" would allow that phrase to encompass all processes which contribute to the production of a natural language expression, E, from some context, C. This leads to a "demonic" picture of generation as illustrated in Figure 1. Certain contexts produced by non-generative processes in a model containing a NLG component trigger that component. The language generator, in addition to producing E, must alter the context sufficiently to inhibit the reproduction of E ad infinitum.

While this picture is sufficiently general to encompass virtually any proposed generative model, it is so non-commital that it does little to explicate NLG. The question is merely resolved into two subissues:

(1) What constitutes an NLG-activating context?

(2) What processes and knowledge are needed to produce an appropriate E in such a context?

Now in fact (1) has not been addressed as a serious problem in most work on generation. The activating context has almost universally been the existence of some "information to be communicated" in a distinguished cell in the context. Any process which "stores" into this cell immediately awakens the generator which proceeds to produce a natural language encoding of that information. Context alteration by the generator consists simply of erasing the special cell.

The paradigm which has evolved out of this decision is depicted in Figure 2.

Models based on this paradigm are differentiated primarily by:

(1) The representations used for messages to be encoded by the generator.
(2) The degree to which the generation box interacts with the context (context-sensitivity of generation).

The predominant formalisms for representing messages are:

(a) (partial) specification of syntactic structure <1>
(b) semantic networks (consisting of case relations between semantic objects) <3,6>
(c) conceptual networks <2>

The dividing line between semantic and conceptual networks is not clear-cut. The intended distinction is that conceptual objects and conceptual relations are divorced from natural language, whereas semantic nets are constructed to represent meaning in terms of objects and relations specifically designed for (some particular) natural language.

Presumably one reason for separating the selection of a message from the task of encoding that message into natural language was to free research on "generation" from the necessity of dealing with context. But in recent years our generation models have become more and more context sensitive - this is true at least of NLG models which treat message encoding as a subpart of some larger task. Some of these contextual considerations appear to be independent of any particular target language - e.g., consultation of context in determining which features of an object should be mentioned in its description <7> - while others depend on detailed knowledge of the target language - e.g., the choice of verbs and nouns to be used in describing events <2>. The increased use of context is done to effect a more "natural" encoding of the message rather than simply a "legal" encoding. In this respect there are implicit in such NLG models certain assumptions about the use of context in language understanding. This matter will be elaborated somewhat later.

The set of processes and knowledge needed to encode a message depends heavily on the message representation chosen. This existence of a formal grammar as the repository of syntactic knowledge about the target language has become standard practice; transition network grammars are representative of the current state of the art in this respect. The progression from syntactic to semantic to conceptual representations entails the use of progressively more knowledge about language and communication. A semantic net representation may need a theory of semantic cases and rules for mapping these into surface cases of the target language; a conceptual representation requires complex rules and extensive data to choose appropriate words for the construction of the target language expression.

III. CONCEPTUAL GENERATION

My own work in NLG falls within the paradigm described above under the assumption that the message to be expressed is available only in a conceptual representation. This means that neither the words of the target language (English) nor the syntactic structures appropriate for encoding the message are initially available to the generator. They must be deduced from the information content of the message. (Actually one exception to this claim is clearly present in the model - an initial presumption is made that the message is encodeable as a single English sentence. This is an unwarrented assumption which hides a potentially significant problem.) Once actual words have been selected and organized into an English-specific syntactic structure, knowledge of English 'linearization' rules - e.g., that adjectives precede the nouns they modify - are used to produce a surface string. This knowledge is contained in an AFSTN grammar and utilized by a method introduced by Simmons and Slocum <6>.

By working from a conceptual representation, a generator assumes the burden of accounting for two aspects of language production generally ignored in other models. The first of these is word selection, which is accounted for by a pattern matching mechanism, namely decision trees (discrimination nets). In order to account for the selection of appropriate words, it is necessary to presume that the generator has extensive access to contextual information as well as access to inferential capabilities and belief structures. The second aspect of generation which must be addressed in the linguistic encoding of conceptual graphs concerns the expression of meaning by structure in addition to its expression by individual words. The case framework of verbs is one source of knowledge which deals with structural encoding - e.g., in English the recipient of an object can be encoded as a syntactic SUBJECT if the verb "receive" is used to describe the transmission of that object. Other forms of structural encoding are not determined by verb-related rules - e.g., that the construction <container> OF <contents> can be used in English to express the relationship between a container and the object(s) it contains.

The generation algorithm demonstrates a mixture of data-driven and goal-driven behavior. In addition to the initial goal - "generate a SENTENCE expressing the meaning of the given graph" - choices made in the course of generation set up sub-goals - e.g., "express the RECIPIENT of a transmission and make it the SUBJECT of the structure being built." The conceptual content of the message, however, drives the selection of a verb for the English sentence

and the construction of 'optional' structural segments.

The choice of conceptual structures as a base for NLG was not made because of any particular designed (or accidental) suitability of conceptual graphs for this purpose. Indeed it would be possible to alter the representations in ways which would simplify the task of generation. But, if a NLG model is to be utilized as a means of transmitting information from a machine to a human, then the construction of that information is a prerequisite of its encoding. More importantly, for uses of generation in "intelligent" systems, the construction of the information is the most time-consuming process. For this reason conceptual structures are designed to facilitate inference and memory integration capacity <5> - if necessary, at the expense of ease of linguistic analysis and generation.

IV. UNDERSTANDING AND GENERATION

For several years there has been a strong emphasis on the problem of understanding in computational models, and relatively little on problem of generation. In the proceedings of the past two International Joint Conferences on Artificial Intelligence, for example, we find eight papers dealing with the analysis of natural language, three describing both analytic and generative portions of language processing systems, and none devoted mainly to NLG. At least two reasons for this bias are discernable:

(1) Resolution of ambiguity, long recognized as one of the central problems of language understanding, relies for its solution on capabilities - limited inference, expectation, hypothesis generation and testing - required by other 'intelligent' behavior. As long as language generation was viewed as basically a matter of codifying linguistic knowledge, it appeared far less relevant to the AI community than did analysis.

(2) For those with a pragmatic bent the lack of symmetry between requirements of an analyzer and those of a generator made research on understanding of paramount importance. That is, for a given domain of discourse, a machine can afford to express information utilizing a limited vocabulary and with limited syntactic variety without placing an unacceptable burden on a human conversant; to ask a human to adhere to equivalent limitations in his own language production could prohibit the conduct of any interactive dialogue (at least without extensive training).

Furthermore, there exist a great many tasks which are currently or will soon be within the capacity of computers and which could be usefully extended by a natural language 'front end' - i.e., an analyzer. Corresponding needs for a natural language 'back end' are harder to find, perhaps because we are so accustomed to using our machines in the computation of numerical or tabular data, which is seldom enhanced by expression in natural language.

Being of a pragmatic bent myself, at least in spirit, I think the bias toward analysis is justified. But I expect that as the boundaries of generation are pushed back and more work is done on the semantic aspects of generation, the view of 'analysis' and 'generation' as disparate endeavors will change considerably. I see far more commonality than disparity in the two enterprises. Both require much the same capacity for inference and deduction, albeit for different purposes. The knowledge of the syntactic structure of a language needed to understand that language is also needed to generate it, although the organization of that knowledge may be different. A similar condition holds for knowledge of word meanings and mappings from syntactic structure to semantic or conceptual relations. Still, I do not believe we are ready for, or should even be striving for, a single representation and organization of this knowledge which would permit its being shared by both analytic and generative processes. But a good deal of the fruits of research can be shared.

V. NEW DIRECTIONS

It seems to me that there exist several problem areas in the development of a complete theory of language generation which have scarcely been touched. Some of these could be, and possibly are being, profitably addressed already; others seem to involve extremely long range goals. Into the latter category I would put the issue of message selection referred to earlier. A theory capable of accounting for message selection in a general context would need a thorough motivational model, probably of both the information producing mechanism and the human information 'consumer'. Fortunately, adequate heuristics for message selection are much easier to develop for task specific domains, so lack of a general theory is not likely to hinder either research on or application of language generation techniques.

However, a great deal can be done in the short range on the use of context in generation: (1) as it relates to the determination of message encoding, and (2) in the modification of context in ways which affect later analysis, generation, and reasoning processes.

Another frontier of research is in the communicative rather than linguistic aspects of NLG. "Message selection" has been used in this paper to refer to the choice of

information to be conveyed to a human. The nature of human communication is such that it is generally necessary only to transmit a subpart of the totality of that message. Context and the understanding mechanisms of the information consumer are capable of filling in much vague or omitted information. Winograd's heuristic for describing toy blocks addresses precisely this issue - it amounts to an implicit model procedurally encoded in a generation program, of a process for finding the intended referent of an object description. While I would not push for incorporating explicit models of understanding in our generation models, I believe much could be gained by the addition of further implicit knowledge of this sort.

BIBLIOGRAPHY

<1> Friedman, J., A COMPUTER MODEL OF TRANSFORMATIONAL GRAMMAR, American Elsevier, New York, 1971

<2> Goldman, N., "Sentence Paraphrasing from a Conceptual Base," CACM Vol. 18, No. 2, February 1975.

<3> Klein, S., et. al., "Automatic Novel Writing: A Status Report," University of Wisconsin TR 186, August 1973.

<4> Riesbeck, C., 'Computational Understanding of Natural Language Using Context', AIM-238, Computer Science Department, Stanford University, Stanford, California.

<5> Schank, R., et. al., "Inference and Paraphrase by Computer", Journal of the ACM, July 1975.

<6> Simmons, R., and Slocum, J., "Generating English Discourse from Semantic Networks", CACM, Vol. 15, No. 10, October 1972.

<7> Winograd, T., "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language", TR-84, M. I. T. Project MAC, February 1971.

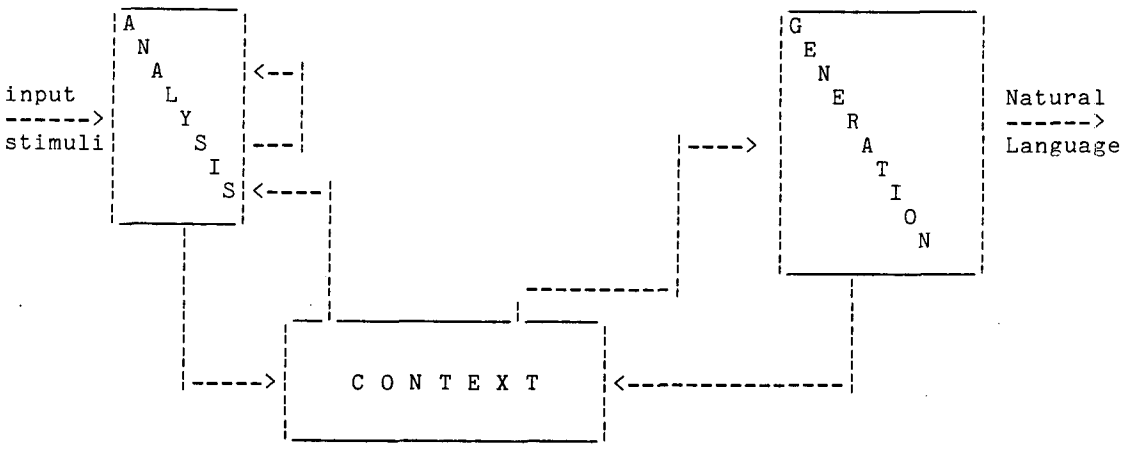

Figure 1

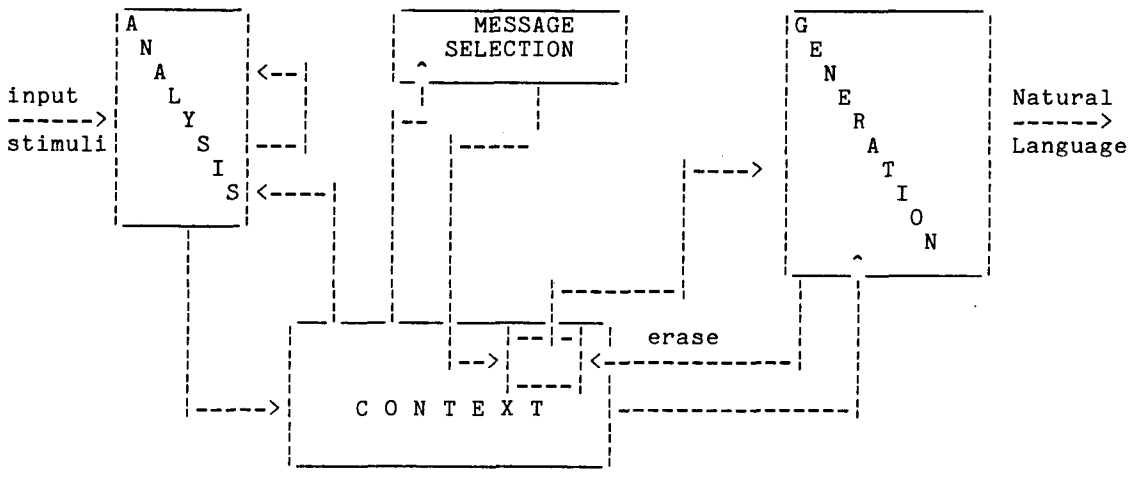

Figure 2