SALIENCE: THE KEY TO THE SELECTION PROBLEM IN NATURAL LANGUAGE GENERATION

E. Jeffrey Conklin
David D. McDonald
Department of Computer and Information Science
University of Massachusetts
Amherst, Massachusetts 01003 USA[1]

ABSTRACT

We argue that in domains where a strong notion of salience can be defined, it can be used to provide: (1) an elegant solution to the selection problem, i.e. the problem of how to decide whether a given fact should or should not be mentioned in the text; and (2) a simple and direct control framework for the entire deep generation process, coordinating proposing, planning, and realization. (Deep generation involves reasoning about conceptual and rhetorical facts, as opposed to the narrowly linguistic reasoning that takes place during realization.) We report on an empirical study of salience in pictures of natural scenes, and its use in a computer program that generates descriptive paragraphs comparable to those produced by people.

I. The Selection Problem

At the heart of research on natural language generation is the question of how to decide what to say and, equally important, what not to say. This is the "selection problem", and it has been approached in various ways in the past: Direct translation generators such as [Swartout 1981, Clancey to appear] avoid the problem by leaving the decision to the original designer of the data structures that serve as the templates to the generator; this places the burden on that designer to correctly anticipate what degree of detail and presupposed knowledge will be appropriate to a specific audience since on-line adjustments are not possible.

Mann and Moore [1981], on the other hand, while assembling texts dynamically to suit their audience, do so by "over-generating" the set of facts that will be related, and then passing them all through a special filter, leaving out those that are judged to be already known to the audience and letting through those that are new. McKeown [1981] uses a similar technique -- her generator, like Mann and Moore's, must examine every potentially mentionable object in the domain data base and make an explicit judgement as to whether to include it. We argue that in a task domain where salience information is available such filters are unnecessary because we can simply define a cut-off salience level below which an object is ignored unless independently required for rhetorical reasons.

The most elaborate and heuristic systems to date use meta-knowledge about the facts in the domain and the listener's knowledge of them to plan utterances to achieve some desired effect. Cohen [1978] used speech-act theory to define a space of possible utterances and the goals they could achieve, which he searched by using backwards chaining. Appelt [1982] uses a compiled form of this search procedure which he encodes using Saccerdotti's procedural nets; he is able to plan the achievement of multiple rhetorical goals by looking for opportunities to "piggyback" additional phrases (sub-plans) into pending plans for utterances. We argue that in domains where salience information is already available, such thorough deliberations are often unnecessary, and that a straight-forward enumeration of the domain objects according to their relative salience, augmented with additional rhetorical and stylistic information on a strictly local basis, is sufficient for the demands of the task.

---

1. This report describes work done in the Department of Computer and Information Science at the University of Massachusetts. It was supported in part by National Science Foundation grant IST#8104984 (Michael Arbib and David McDonald, Co-Principal Investigators).

II. Deep Generation and Scene Descriptions

In this paper we present an approach to deep generation that uses the relative salience of the objects in the source data base to control the order and detail of their presentation in the text. We follow the usual view that natural language generation is divided into two interleaved phases: one in which selection takes place reflecting the speaker's goals, and the selected material is composed into a (largely conceptual) "realization specification"[1] (abbreviated "r-spec") according to high-level rhetorical and stylistic conventions, and a second in which the r-spec is realized -- the text actually produced -- in accordance with the syntactic and morphological rules of the language. We call the first phase "deep generation" -- instead of the more specific term "planning" -- to reflect our view that its use of actual planning techniques will be limited when compared to their use in the generators developed by Cohen, Appelt, or Mann and Moore.

We are developing our theory of deep generation in the context of a computer program that produces simple paragraphs describing photographs of natural scenes similar to those analyzed by the UMass VISIONS System [Hanson and Riseman 1978, Parma 1980]. Our input is a mock-up of their final analysis of the scene, including a mock-up annotation of the salience of all of the objects and their properties as would be identified by VISIONS; this representation is expressed in a locally developed version of KL-ONE. The paragraphs are realized using MUMBLE [McDonald 1981, 1982], which is responsible for all low-level linguistic decisions and for carrying out the rhetorical directives given in the r-spec.

1. We are introducing this new term -- "realization specification" -- in place of the term "message" which had been used in earlier publications on McDonald's generation system. This is a change in name only; these objects have the same formal properties as before. The shift reflects the kind of communication metaphor on which this work has actually been based: the old term has often connoted a view of communication as a process of translating a data structure in the speaker's head into language and then reconstructing it in the audience's head. (the so-called "conduit" metaphor). Instead, we take it that a speaker has a set of goals whose realization may entail entirely different utterances depending upon who the audience is and what they already know; that the speaker's knowledge of their language consists in large part of a catalog of what might be said and the effects it is likely to have on the audience; and that, accordingly, language generation entails a planning process, selecting among these effects according to the desired outcome.

As of the beginning of February 1982, the initial version of the deep generation phase has been designed and implemented. Figure 1 shows the kind of scene we are using in our studies and an example of the kind of paragraph description targeted for our system. Efforts to

"This is a picture of a large white house with a white fence in front of it. In front of the fence is a cement sculpture. In front of this is a street. Across the street is a grassy patch with a white mailbox. There are trees all around, with one evergreen to the right of the driveway, which runs next to the house. It is fall, the sky is overcast, and the ground is wet."

Figure 1. One of the pictures used in the experimental studies with one of the subjects' descriptions of it. A mocked-up analysis of this picture was used as the input to the deep generation process in the example discussed below.

modify MUMBLE to run in NIL on our VAX are underway, and we anticipate having an initial realization dictionary up and the first texts produced before the end of May. During the summer and fall of 1981, Jeff Conklin (Conklin and Ehrlich, in preparation) carried out the series of psychological experiments discussed immediately below. The results have been used to determine the salience ratings for the mock-up of the analyzed scenes, and to provide a corpus of the kinds of texts people actually produce as descriptions of scenes of suburban houses.

III. Visual Salience

Our theory of visual salience states that a given person looking at a given picture in a given context assigns a salience (an ordering, rather than a numeric value) to each object as a

natural and automatic part of the process of perceiving and organizing the scene. Intuitively the salience of an object is based on its size and centrality (how central it is) in the image, its degree of unexpectedness, and its intrinsic appeal or importance to the viewer.

To substantiate and explore these intuitions we ran a series of experiments in which a group of subjects rated the salience of items in color slides of natural scenes. For each picture each subject had a form listing all of the major items in the scene, and their task was to rate the salience of each item on a zero to seven scale. In order to define a controlled context the subjects were asked to imagine that they worked for a library which had a large picture section, and that their ranking scores would be used to catalog the pictures. The controlled context is necessary because salience is generally only defined within a perceptual or conceptual context -- there is no salience in a vacuum. (However, we claim that there is a default context for viewing pictures which "anchors" the notion of salience when no other context is specified: that pictures are taken for the purpose of showing or telling the viewer something. While this is not a strong context, it allows one to talk about visual salience without precisely defining a purpose for the viewer.)

In several experiments the subjects were given a second task: writing a description of the same pictures for which they were doing the rating task (one such description appears in Figure 1). In these experiments the series of pictures was shown twice; in the first viewing, half of the subjects did the rating task and the other half did the description task, while in the second viewing the tasks were reversed. (It turned out that the description task had no significant effect on the rating scores.)

Although we are still analyzing the data from these experiments, there are several interesting results. The rating technique is a fairly stable and consistent non-subjective measure of salience (when averaging over a group), and is also quite sensitive to changes in the size and centrality of objects in the scene. Figure 2 shows a series of pictures that were used to determine the affects of size and centrality. The salience ratings assigned by subjects to the parking meter in this series were significantly different from each other ($P<.05$, as measured by the Wilcoxon rank sum test). That is, the rating task is sensitive enough to reveal small changes in the size and/or centrality of objects in a picture.

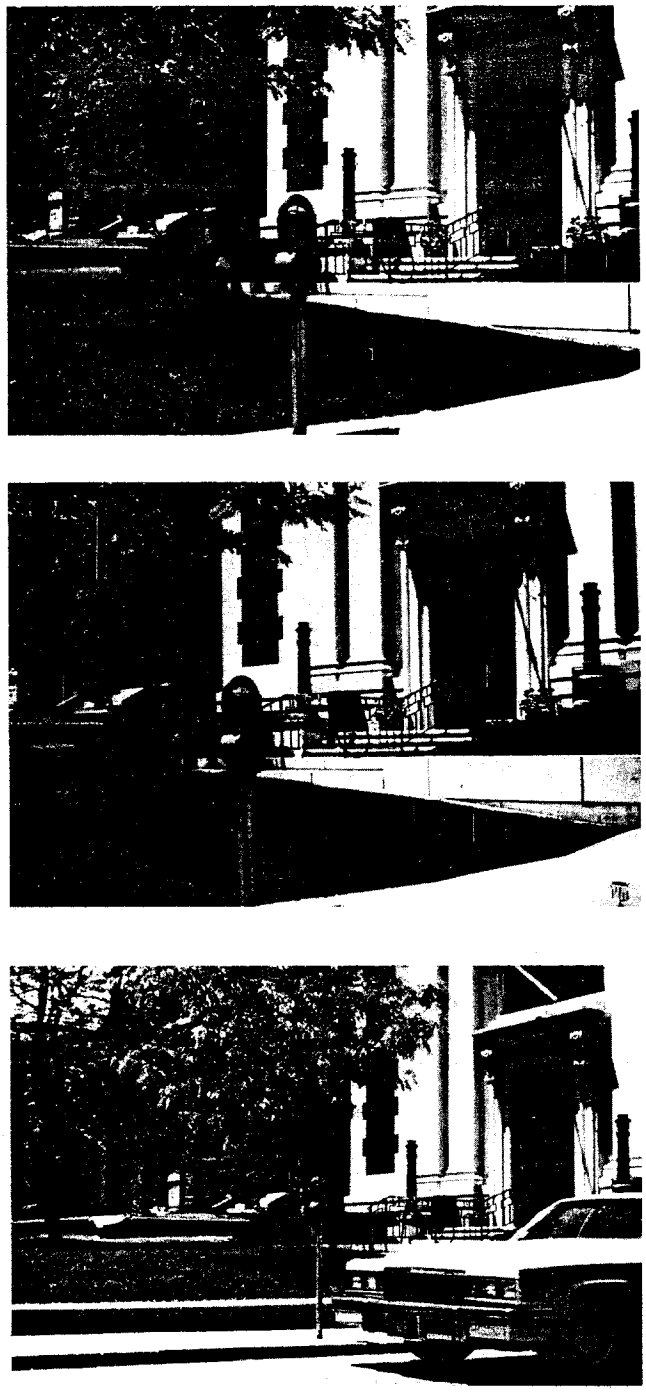

Figure 2 A series of views of a parking meter used to measure the affects of size and centrality.

Also, it was found that salience was a strong determinant in the order of mention of objects in the paragraphs. Specifically, the higher the salience rating given an object by a subject, the more likely that object was to appear in the subject's description. Furthermore, there was a good correlation between the ranking of the objects (by decreasing salience) and the order in which the objects were mentioned in the description. Interestingly, the exceptions to a perfect correlation were generally the cases where a low salience item was "pulled up" into an earlier position in the text, seemingly for rhetorical reasons. The explanation that we propose is that salience is the primary force in selection in scene descriptions, but that rhetorical factors can override it (as illustrated below).

IV. An Example

Here is an short example of the kind of paragraph which our system currently generates:

> "This is a picture of a white house with a fence in front of it. The house has a red door and the fence has a red gate. Next to the house is a driveway. In the foreground is a mailbox. It is a cloudy winter day."

This paragraph was generated from a perceptual representation (in KL-ONE) in which the most salient objects, in order of decreasing salience, were:

House, Fence, Door, Driveway, Gate, and Mailbox.

The deep generation component (called GENARO) maintains this list as the "Unmentioned Salient Objects List" (USOL), and it is this data structure which mediates between GENARO and the domain data base (see Figure 3). It should be stressed that the USOL contains only objects -- not properties of objects or relationships between objects -- since we specifically claim that such an "object-driven" approach is not only more natural but also is adequate to the task.

There are two "registers" which are used for focus: "Current-Item" and "Main-Item". The Current-Item register contains the object currently in focus (and hence the most salient object which has not previously been mentioned), and the Main-Item register points to the data base's most salient object as the topic of the entire paragraph (this register is set once at the beginning of the paragraph generation process). An object moves into focus by being "popped" from the USOL and placed in the

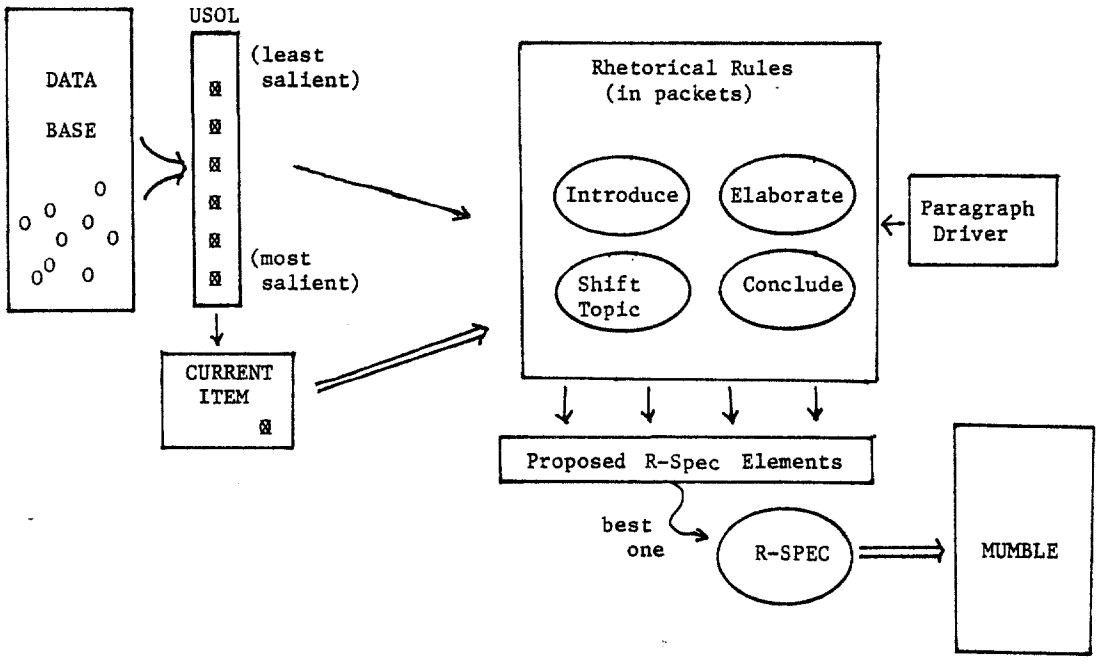

Figure 3. A block diagram of the GENARO system. The "O"s in the "Data Base" represent objects in the domain representation, whereas the "⊠"s are the themeatic "shadows" of these objects used by GENARO for its rhetorical processing. Each of the ovals in the "Rhetorical Rules" box are packets containing one or more rhetorical rules.

Current-Item register, along with its most salient properties and relationships (for ease of access). When formulating the r-spec, most of the rhetorical rules then look only at the Current-Item. (Some rules look down "into" the USOL, or into the r-spec under construction, as elaborated below.)

GENARO stores its rhetorical conventions in the form of production rules, which are organized in packets (a la Marcus, 1980). The packets are used for high-level rhetorical control (i.e. introducing, elaborating, shifting-topic, concluding), and are turned on and off by a Paragraph Driver (which encodes the format of descriptive paragraphs). We call this control structure for the production rules "Iterative Proposing": each of the rules in the active packets whose condition is satisfied makes a proposal and gives it a rhetorical priority; the proposals are then ranked, and the one with the highest priority wins. This process is itterated until the r-spec is complete. The environment in which the rules' conditions are evaluated may change from itteration to itteration as a result of actions performed by the winning proposals. The r-spec can thus be thought of as a "molecule", each of whose "atoms" is the result of a successful rule. The atoms are "specification elements" to be processed by MUMBLE; they are either objects, properties, or relations from the domain, or rhetorical instructions that originate with GENARO. (N.b. In the course of producing a paragraph many r-specs will pass from GENARO to MUMBLE. The flow of the paragraph is determined by which rules are turned on -- via the Paragraph Driver's control of which packets are on -- and each r-spec is produced "locally", without an awareness of previous r-specs or a planning of future ones.)

GENARO starts with an empty message buffer and with Current-item (in our example) set to House, the first item in the Unused Salient Object List. The Introduce packet, which is turned on initially, has a rule which proposes to "Introduce(House)"; this rule's conditions are that the value of the Current-Item be value of the Main-Item (i.e. the Main-Item is in focus), and that the salience of the Main-Item be above some specified threshold. In this example both of these conditions are met, and the "atom" Introduce(House) is proposed at a high rhetorical priority, thus guaranteeing not only that it will be included in the first r-spec, but that it will be the dominant atom in that r-spec. Another rule (in the Elaborate packet), proposes including the color of the house (e.g. Color(House,White)), not because the color is itself salient, but to "flesh out" the introductory sentence. This rule is included because we noticed that salient items were rarely mentioned as "bare" objects -- some property was always given. (Note also that there are other rules that propose mentioning properties of objects on other grounds, i.e. because the property itself is salient.) Finally, there is a rule which notices that Fence is both quite salient and directly related to the current topic, and so proposes In-Front-Of(Fence, House).

Since the r-spec now contains three atoms and there are no strong grounds based on salience or considerations of style to continue adding to it, the r-spec is sent (via a narrow bandwith system message) to the process MUMBLE, which immediately starts realizing it. MUMBLE's dictionary contains entries for all of the symbols used in the r-spec, e.g. Introduce, In-front-of, House, etc., which are used to construct a linguistic phrase marker which then controls the realization process, outputing "This is a picture of a white house with a fence in front of it.". Back in GENARO, after the r-spec was sent, the Introduce packet was turned off, the message buffer cleared, Door (the next unused object) removed from the USOL and placed in the Current-Item register, and the Iterative Proposing process started over.

In building the next r-spec, Part-of(Door, House) and Color(Door, Red) are inserted, by rules similiar to the ones described above. Suppose, however, that there are no other salient relations or properties to mention about the Current-Item Door: nothing of high rhetorical priority is left to be proposed (n.b. once a rule's proposal is accepted that rule turns itself off until that r-spec is complete). There is, however, a rule called "Condense" which looks for rhetorical parallels and proposes them at low priority (i.e. they only win when there are no, more useful, rhetorical effects which apply). Condense notices that both Door (the Current-Item) and Gate (which is somewhere "down" in the USOL) have the property Red, and that the salience of Gate and of the property Color(Gate, Red) are above the appropriate thresholds, and so proposes that Gate be made the local focus. When this action

is taken, a conjunction marker is added to the r-spec, and Gate is pulled out of the USOL and made the Current-item. The r-spec created by these actions is realized as "The house has a red door and the fence has a red gate.".

When the USOL is empty the Conclude packet is turned on, and a rule in it proposes the r-spec about the lighting in the picture. (The facts about "cloudy" and "winter" are present in the perceptual representation -- no extra generation work was done to make that message.)

V. A Rhetorical Problem

One of the issues that we are using GENARO to investigate is that in their written descriptions people sometimes "chain" spatially through a picture, linking objects which are spatially close to each other or are in certain other strong relationships to each other. The paragraph in Figure 1 contains a good example of this style -- the rhetorical skeleton is:

This is a picture of an A with
  a B in front of it.
In front of the B is a C.
In front of the C is a D.
Across the D is an E.

As can be seen by inspecting the picture in Figure 1, A thru E (i.e. house, fence, sculpture, street, and grassy patch) are arrayed from background to foreground in the picture in a way which allows the "in-front-of" relation to be used between them.[1] The question is: By what mechanism do we allow the strong spatial links between these items to override the system's basic strategy of mentioning objects in the order of decreasing salience?

The first part of the answer is that the machinery for such chaining already exists in the way the Current-Item register is used (and can be reset) by the rhetorical rules. Since one of the actions rules are allowed is to reset the Current-Item to some object, a rule can be written which says "If the Current-Item has a salient relationship Relation to object X, then propose Relation(Current-Item,X) and make X the Current-Item". This rule (let's call it Chain) would have the effect of chaining from object to object as long as no other rules had a higher (rhetorical) priority and the various "Relation"'s of the respective Current-Items were salient enough to satisfy the rule's condition.

But this kind of chaining would only happen as the result of a happy series of the right local decisions -- each successful firing of Chain would be independent of the others. Furthermore, there would be no guarantee that the successive "Relation"'s would be the same, as is the case in the above example. What is needed, perhaps, is to give Chain the ability to look at the structure of the evolving r-spec and to notice when there is an opportunity to build upon a structural parallel (e.g. X in front of Y, Y in front of Z). We are currently investigating ways to make this kind of structural parallel visible within r-specs and still maintain them as a concise and narrow-bandwidth channel between GENARO and MUMBLE.

1. "Across" in this case would be a lexical variation on "in-front-of" introduced deliberately by MUMBLE to break up the repetition.

VI. References

Appelt, D. Planning Natural Language Utterances to Satisfy Multiple Goals, Ph.D. Disseration, Stanford University, to appear as a technical report from SRI International, 1982.

Clancey, W. (to appear) "The Epistemology of a Rule-Based Expert System: A Framework for Explanation", Journal of Artificial Intelligence; also available as Heuristic Programming Project Report 81-17, Stanford University, November 1981.

Cohen, P., On Knowing What to Say: Planning Speech Acts, University of Toronto, Technical Report 118, 1978.

Conklin E. J. (in preparation) PhD. Dissertation, COINS, University of Massachusetts, Amherst, 01003.

----- and Ehrlich K. (in preparation) "An Investigation of Visual Salience", Technical Report, COINS, U. Mass., Amherst, Ma. 01003.

Hanson, A. R. and Riseman, E. M. "VISIONS: A Computer System for Interpreting Scenes", in Computer Vision Systems, Hanson, A. R. and Riseman, E. M. (Eds), Academic Press, New York, pp 449-510, 1978.

Marcus, M. A Theory of Syntactic Recognition for Natural Language, MIT Press, Cambridge, Massachusetts, 1980.

McDonald, David D. "Language Generation: the source of the dictionary", in the Proceedings of the Annual Conference of the Association for Computational Linguistics, Stanford University, June, 1981.

----- "Natural Language Generation as a Computational Problem: an introduction" in Brady ed. "Computational Theories of Discourse", MIT Press, to appear, fall 1982.

McKeown, K. , Generating Natural Language: Deciding What to Say Next, University of Pennsylvania, Technical Reprot MS-CIS-81-1, 1981.

Mann, W. and Moore, J. "Computer Generation of Multiparagraph Text", American Journal of Computational Linguistics, 7:1, Jan-Mar 1981, pp 17-29, 1981.

Parma, Cesare C., Hanson, A. R., and Riseman, E. M. "Experiments in Schema-Driven Interpretation of a Natural Scene", in Digital Image Processing, Simon, J. C. and Haralick, R. M. (Eds), D. Reidel Publishing Co., Dordrecht, Holland, pp 303-334, 1980.

Swartout, W. Producing Explanations and Justifications of Expert Consulting Programs, Technical Report 251, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1981.