SALIENCE AS A SIMPLIFYING METAPHOR FOR NATURAL LANGUAGE GENERATION

David D. McDonald and E. Jeffery Conklin
Department of Computer and Information Science
University of Massachusetts
Amherst, Massachusetts 01002 USA[1]

Abstract

We have developed a simple yet effective technique for planning the generation of natural language texts that describe photographs of natural scenes as processed by the UMass VISIONS system. The texts follow the ordering on the scene's objects that is imposed by their visual salience -- an ordering which we believe is naturally computed as a by-product of visual processing, and thus is available -- for free -- as the basis for generating simple but effective texts without requiring the complex planning machinery often applied in generation. We suggest that it should be possible to find structural analogs to visual salience in other domains and to build comparably simple generation schemes based on them. We look briefly at how one such analogy might be drawn for the task of tutoring novice PASCAL programmers.

I Natural language generation and the superhuman-human fallacy

Taken in its general form, the problem of deciding what to say is a planning problem of great complexity. When speaking carefully and deliberately a person will attempt to simultaneously satisfy many goals from different sources: rhetorical, tutorial, affective, and descriptive, among others. Utterances are intended to obey strict constraints deriving from the limited expressive power of the syntax and vocabulary of natural language and from the requirement to maintain the linguistic coherency of the discourse context established by what has been said up to that point. In addition utterances must do all this while being reasonably short in length and precise in style if the audience is not to become bored or confused. It is no wonder, then, that the ability to speak or write well does not come easily. Even though we all use language constantly, relatively few of us have the skill of a Mark Twain or a Winston Churchill. The requirements of everyday communication do not appear to require optimum linguistic performance.

In this light, we must consider whether we have been making the generation probelm for computers more difficult than it actually is for people -- the superhuman-human fallacy. Should we require our computers to speak any more effectively than we do ourselves? Most of us, as we speak, notice when we have left something out or inadvertently given the wrong emphasis, and we correct our mistakes by interrupting or modifying what we were about to say next; in explanations we use feedback from our audience such as questions or puzzled looks to dynamically adjust our vocabulary and level of detail. We should seriously consider designing our natural language generation systems on a similar basis: adopting an expedient and computationally efficient, if "leaky", planning process and compensating for it by monitoring and attending to user questions.

At the University of Massachusetts we have developed just such an expedient planning system, which we use in conjunction with a highly efficient (i.e. quasi-realtime) text generator. Taking as input a simulation of the output of a computer vision system, the planner determines the order in which objects will be mentioned in the text and what will be said about them, feeding this information via a pipeline to the generator where grammatical constraints determine the exact phrasing and local rules (such as pronominalization and ellipsis) are applied to maintain the coherency of the discourse.

The key to the planner's simplicity is its reliance on the notion of "salience" -- objects are introduced into the text according to their relative importance in the conceptual source of the text. The decision as to what objects, properties, and relations to leave out -- a source of considerable labor in some generation systems (e.g. Mann and Moore [6], McKeown [5]) -- is handled trivially here by defining a cut-off salience rating below which objects are ignored. The task for which we developed this facility, the production of short paragraphs describing photographs of houses, is deliberately one in which the common sense notion of visual salience is vivid and widely shared by members of this culture. People interpret what is important about a picture -- what it is a picture "of" -- according to a shared set of conventions involving the size and centrality of the objects shown, coupled with a sense of what is normal or expected: a large stained-glass window on an otherwise ordinary New England farm house would be highly salient;

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).

similarly a normally unimportant part of the scene, such as the mailbox, can be artificially raised in salience if framed prominently in the foreground of the picture.

II Our Generation System

As of this writing, the salience-based planner (the subject of Conklin's PhD. thesis) has been implemented and its pipeline to the text generator (McDonald's system MUMBLE [9]) hand simulated. The house scenes which are the source of the text are very similar to those used in the research of the UMass "VISIONS" system [10] (see Figure 1); their representation is also presently hand-simulated: the planner works from a KL-ONE data base of the objects in the scene and the spatial relations between them which was designed in close collaboration with members of the VISIONS project, and which reflects the actual kinds of information they expect to extract from a visual scene.

The salience ratings with which the objects in the visual representation are annotated were derived empirically through extensive psychological testing of human subjects [3], where the subjects both rated the objects in each of a series of pictures on a zero to seven scale, and wrote short paragraphs describing the scenes. The objects' ratings were quite consistent across subjects and sessions of the experiment. The paragraphs provide an objective base-line for the kind of style and overall organization that should be generated by the system.

Fig. 1.
One of the pictures used in the studies and an example of the kind of descriptive paragraph that subjects wrote about it.

"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. There is a driveway beside the house, and a tree next to the driveway. In the foreground is a mailbox. It is a cloudy day in winter."

Given the salience data, the planning algorithm runs as follows (see also [4]): the objects in the scene are placed in a list -- the "Unused Salient Object List" -- in decreasing order from most to least salient. The properties of the objects (such as color, size, or style) and their relative spatial relations can be accessed from the general scene data base when desired; one can, for example, ask for the most salient relationship in which a particular object is involved (by definition relations acquire their salience from the objects they relate). Objects are taken from the "Unused Salient Object List" (shortening the list in the process), packaged with selected properties and relations, and sent to the generator by the action of a collection of strictly local rhetorical rules. The rules are couched as productions, have relative priorities, and are organized into packets according to when they apply -- essentially the same architecture as Marcus used in his natural language parser [7]. This architecture allows us to incorporate object-specific rules (such as that one always sees houses introduced with one of their properties: "a white house" or "a New England farm house", and never simply as "a house") and also simple stylistic rules, such as maintaining sentences of an appropriate length.

The process proceeds by successively taking the first object on the list (i.e. the most salient unmentioned object), making it the local "current item", and proceeding to describe the most salient properties and relations, finally "popping" the list of unmentioned objects and moving on to describe the next most salient object.

The scene descriptions produced by this process will never win a prize for good literature. They are, however, apparently effective as descriptions: as judged by (so far only a few) informal trials, paragraphs generated automatically on the basis of the salience ratings derived from the experiments are effective in picking out which picture they correspond to from others of the same material but taken from a different camera angle. Furthermore they provide a base line for potentially measuring the "value-added" of a full-scale global planning system that would be capable of reasoning about and directing the larger-scale rhetorical organization of the text (say, one on the model of Appelt [1], or McKeown [5]).

III Where does salience come from?

We claim that the annotation of an object's visual salience can be provided as a natural part of the perception process. For example, one aspect of salience stems from unexpectedness: items which are not predicted by, or are inconsistent with, high-level world knowledge are unusual and therefore salient. Also, an item's size and centrality in the picture are clearly factors in that item's salience. Specifically, the record of an object's relative salience would arise from the perceptual process's explicitly combining: 1) the weighting the object contributed to the judgement that the scene was what is was, 2) the object's

intrinsic salience (from general world knowledge, e.g. people are intrinsically more salient than bushes in conventional pictures), and 3) the amount of "extra effort" that was required to resolve the assignment of the object to a slot of the frame when default assumptions were violated. The salience annotation of the visual representation is consequently provided as a direct part of the perceptual analysis of the picture, or is no more than minimal additional on-line computation. (Moreover, the perceptual analysis is the only stage at which salience values can be reasonably cheaply computed.) As a result, a salience-based planner consumes less overall computational effort than a more conventional language planner -- the salience information is provided at no additional computational cost by an already present, non-linguistic process, and it acts as a powerful heuristic in guiding the process of deciding what to say and what to leave out.

While the common sense concept of "salience" applies most naturally to perceptual domains, if we are to draw on salience as an organizing metaphor for language generation in non-perceptual domains, such as tutoring or explanations of physical processes, then we must step back from the specific measures listed above. What role does salience play in the coordination of our thinking about a picture? What kinds of computational processes does it reflect?

To answer these questions we must cast the "goals" of the visual process in terms that carry over into non-perceptual domains. The approach of the VISIONS system is to combine bottom-up analysis of regions and edges in the raw visual image with top-down testing of frame-based hypotheses about the subject matter of scene. The VISIONS system is thus model-driven once it moves beyond the low-level vision problem of identifying regions. For example, once the system has enough edge and region data to suggest that there is a house in the image it will apply its generic knowledge about what parts houses typically have and how they are typically spatially arrayed to attempt to impose identifications on regions which would otherwise be ambiguous. Note that even if the image is actually a picture of, say, a boat in the water, it is still possible that elements of the boat's image might initially trigger the house hypthesis; in this case elements of the picture which were inconsistent with the house scene frame, such as the blue color of the ground plane, would be vital in cutting off expensive attempts to further instantiate that frame.

Broadly speaking, the process of perception can be viewed as a process of building an internal model of some external "world" based on "sensory" data from that world and generic knowledge about it. In this light the components of salience can be described more abstractly. First, the system relies on the conventions of centrality and size of a region to direct its attention so that its first analyses are of those parts of the photograph which are most likely to yield a potent model for identifying the rest of the scene. Second, elements of the image which are unexpected (i.e. which do not have a good "fit" with their slot in the hypothesized frame) are important to the efficient allocation of resources, and would likely be annotated with some measure of their "goodness of fit". Finally, information about the intrinsic importance of various items in the scene might be useful in the allocation of additional resources to the confirmation of their indentification (e.g., if the system is told, as part of its world knowledge, that people are intrinsically important, it would want to be especially sure when it identified image regions as people).

To summarize, these are the elements of model building for which the notion of salience is especially important: 1) structural knowledge about where in the external field of data to focus resources initially (e.g. size and centrality); 2) use of a measure of "goodness of fit" to direct the competition and instantiation of generic frames; and 3) a priori knowledge about what "objects", if found in the "world", are particularly important to the system (i.e. intrinsic salience).

IV Salience in a tutoring task

We are beginning to see this confluence of model-building knowledge sources plus deviation from defaults as perhaps the source of saliency in another domain where we are working on natural language generation: the planning of tutorial dialogues.

Beverly Woolf, an advanced graduate student working with McDonald, is extending the work of the MENO-II project [11] on identifying conceptual misconceptions in simple loop programs, so as to develop a program capable of tutoring the student on a selected subclass of these misconceptions. Analogous to the parameters of size and centrality, the MENO-II project has knowledge about PASCAL errors and their relationships to various misconceptions: this is the starting point for the tutor's analysis. Analogous to the VISIONS system's generic knowledge about possible objects and scenes, the MENO-II project has a very rich knowledge base of the plans and schemas inherent in correct loop programs and their relationship to the student's coherent (but inappropriate to PASCAL) model of the algorithms; a rich, hierarchically organized KL-ONE representation is used for this purpose, including a taxonomy of common misconceptions. Finally, analogous to VISION's a priori knowledge about intrinsic importance, the tutoring system has certain "bugs" and misconceptions flagged as especially important and revealing if found in the model of the student.

The tutoring program is still in the early design stages, consequently we cannot yet be certain that our strategy of applying the salience metaphor to this problem will succeed. However, our intention is as follows: the model of the history of the student will be projected onto the general knowledge base of program techniques and typical misconceptions, where it will be used to identify those parts of the knowledge base which are most relevant to the student's problem, and to suggest a tutoring strategy which will build most

effectively on what the student already knows and will carry them forward to the most appropriate programming concepts for them to learn next. Said another way, the student's buggy program will typically contain several misconceptions on which the student could be tutored. The general knowledge base (interpreted now in terms of plausible tutoring strategies, i.e. alternative sequences of examples and probing questions) will provide a backdrop on which to project the student's specific history so as to pick out the best strategy for that case. By monitoring this analysis process we should be able to annotate specific concepts in the knowledge base and points in the buggy program with a rating analogous to that of visual salience: this annotation will yield the ordering in which those points and concepts are to be taken up during the tutoring session.

References

[1] 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).

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

[3] ----- and Ehrlich K. (in preparation) "An Empirical Investigation of Visual Salience and its Use in Natural Language Processing", Technical Report, COINS, U. Mass., Amherst, Ma. 01003.

[4] ----- and McDonald "Salience: The Key to the Selection Problem in Natural Language Generation", in the Proceedings of the Association for Computational Linguistics, Toronto, Canada, 1982.

[5] McKeown, K. R. Generating Natural Language Text in Response to Questions about the Data Base Structure, Ph.D. Dissertation, Moore School of Electrical Engineering, University of Pennsylvania, 1982.

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

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

[8] 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.

[9] -----, Natural Language Generation as a Process of Decision Making under Constraint, PhD. dissertation, MIT, 1980.

[10] 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, 1980 pp 303-334.

[11] Soloway, Elliot, Beverly Woolf, Eric Rubin, and Paul Barth "Meno-II: An Intelligent Tutoring System for Novice Programmers", Proceedings of International Joint Conference in Artificial Intelligence, Vancouver, British Columbia, 1981.