

CONTENT SELECTION AND ORGANIZATION AS A PROCESS INVOLVING COMPROMISES

Helmut Horacek
Universität Bielefeld, Fakultät für Linguistik und Literaturwissenschaft
Universitätsstr. 25, 33615 Bielefeld, Deutschland

ABSTRACT

Understanding a fairly complex message may demand an increasing degree of effort on behalf of the addressee, a fact that has been neglected almost completely in automated approaches to natural language generation so far. Encountering this problem requires making compromises by reducing the degree of detail in which information is presented, or by explicitly expressing information left implicit otherwise. We identify factors that influence the comprehension effort, and we develop a model that indicates a rough quantified estimate for this effort. We illustrate these ideas by examples of explanations comprising a considerable number of arguments, and we discuss potential impacts of these measurements on the process of composing a message by making compromises.

1. MOTIVATION

Communication in natural language may occasionally fail if the effort associated with mentally capturing the information conveyed becomes too high. In order to avoid this problem, generation programs must have moderately accurate estimates at their disposal, that capture the concepts 'degree of explicitness/implicitness' and 'amount of information/degree of detail' to give an indication of the effort associated with understanding the messages they create¹.

We illustrate these considerations by discussing explanations that illustrate proposals made by an expert system. The system assigns employees to rooms, thereby meeting requirements to provide resources needed, to avoid social conflicts, etc. These requirements break down into constraints, which are relevant for the problem solving process, and into justifications, which give the domain-specific rationale behind these constraints. When the system comes up with a set of assignments for employees (here: A, B, C, and D) to rooms (here: 1, 2, and 3), the user may ask questions about hypothetical assignments deviating from the proposed solution, expecting the system to find out a relevant part of the problem specifications which are responsible for the infeasibility of the assignments focussed on. Figure 1 represents the text plan of an explanation telling why C and D must be in room 3. The plan breaks down into two subplans addressing rooms 1 and 2, respectively, each of which consists of arguments and justifications.

¹ We believe that these considerations apply to oral as well as to written presentations, since being forced to reread portions of a text frequently is hardly desirable. A similar problem to this issue lies in adequately determining the node size in a hypertext document.

There are several possibilities to express this text plan (or some portion of it) in terms of natural language texts. The simplest version (only comprising propositions 1 and 5) is

1) Room 1 must be assigned to A, and room 2 must be assigned to B.

which is easily comprehensible, but hardly informative. Hence, more details should be added, yielding, for instance

2) Room 1 must be assigned to A because A is a group leader, and room 2 must be assigned to B because B is a smoker (additionally comprising propositions 3 and 5).

3) Room 1 must be assigned to A because A as a group leader must be in a big single room and room 1 is the only big single room available. Room 2 must be assigned to B because B must not share a room with either C or D, since B is a smoker, and room 2 is the last single room out of two rooms left (comprising propositions 1 to 3).

Neither of these explanations seems to be entirely satisfactory. Whereas text 2) leaves too much burden on the addressee's inferential capabilities, text 3) is simply too long. Much is left implicit in text 2), which might be recoverable in principle, provided the addressee is well acquainted with the domain regularities (for instance, that group leaders must be assigned to big single rooms), but it is doubtful whether the addressee will really succeed in doing so. In text 3), we suspect, the addressee might easily forget some part of the argumentation before the explanation is completed. Hence, some compromise is required, like:

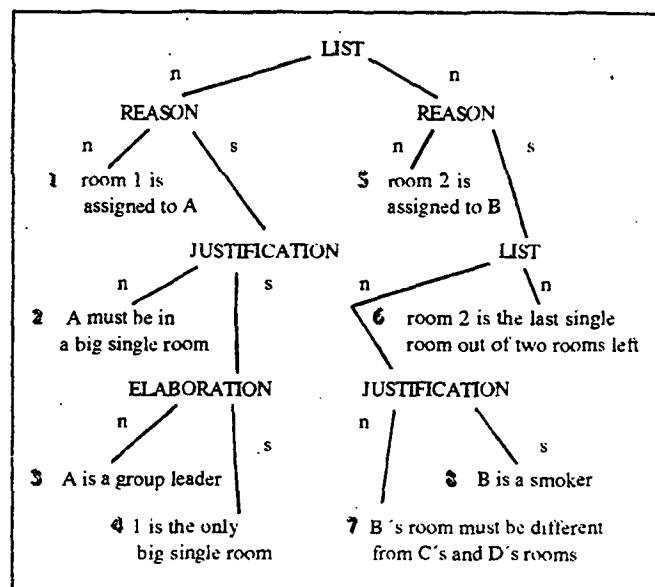


Figure 1: A text structure including arguments in full detail

- 4) Room 1 must be assigned to A because A must be in a big single room, and room 2 must be assigned to B because B must not share a room with either C or D, and because room 2 is the last single room out of two rooms left (comprising propositions 1, 2, 3, 6 and 7, cutting off the justifications).
- 5) Room 1 must be assigned to A because A as a group leader must be in a big single room and room 1 is the only big single room available. And room 2 must be assigned to B (comprising propositions 1 to 4, and elaborating on 1).

We believe that variants 4) and 5) are fairly comprehensible and provide some useful information, but not all that is available. However, if the explanation seeking person is interested in further detail, he/she may ask for justifications to text 4), and for elaborations of text 5).

In the following, we briefly review the contributions to the concepts 'effort required to understand a message' and 'degree of detail entailed in a message' made in the field so far. We describe a first sketch of a formal model that captures these two concepts and accounts for the compensative effect between them. We discuss examples which illustrate the tension between informativeness and comprehensibility including possible compromises between these two factors.

2. PREVIOUS APPROACHES

There is ample evidence from psychological experiments that humans draw causal inferences during reading to close gaps left implicit in narrative texts ([12]). This reasoning is presumably done by building forward-oriented expectations and by drawing backward-driven inferences [5]. In assessing the processing effort associated with understanding the relation between two subsequent sentences, [18] distinguish between "direct causes", which can easily be understood, and "indirect causes" and "unrelated facts", which both require considerable reasoning effort - to infer or, at least, guess a relation or to resign in the attempt of doing so. The distinction between "direct" and "indirect causes" has been made on a statistical basis, which expresses commonalities about default expectations holding across a set of subjects².

In automated approaches to natural language generation, the addressee's comprehension process is anticipated in a few aspects only. They comprise: the avoidance of a potential ambiguity caused by a particular referring expression [10], the inferability of additional information to convey from those propositions uttered explicitly [7], and the proper use of basic level categories like 'dog' and 'house' to avoid false implicatures [16]. However, the mechanisms developed are mainly motivated by detecting potential sources of misinterpretations and by avoiding the production of redundant text. In addition, active checking of the user's understanding has been incorporated in a system concerned with providing explanations [2], and a selection mechanism is used in a reactive approach by Moore and Swartout [15], which incor-

² Within the cases classified as "indirect causes", some expectations triggered by the fact presented in first place constituted "direct causes" of the fact presented in second place, while the "direct causes" of some facts presented in second place were too specific to be expectations triggered by the facts presented in first place.

porates decisions about the most promising explanation strategy under the context of a given situation.

Some series of experiments have been carried out to obtain empirical evidence about how much information humans can typically understand, and what can be considered too much to be memorized without problems. Efficiency in language processing is then influenced to a great extent by finding an optimal mixture of old and new information according to the working memory's capacity [11], which is a capability that may vary strongly among individuals hence, it is almost impossible to identify precise quantifications on general grounds. The only relevant experiments we know of have been reported on a long time ago in [13], which have resulted in the concept of the magical number *seven plus/minus two*. The major reason for this situation is that the variety of knowledge that helps humans in building conceptual chunks is enormous, and this knowledge causes significant differences in their memory capabilities.

Very few systems developed in the field have the capability of producing messages in significantly varying degrees of detail when starting from comparable content specifications. Hierarchical organization is exploited for content selection (in BLAH [19] and in EPICURE [4]). The hierarchy is cut off at some level, eventually due to the user's assumed domain expertise, and only more abstract specifications are included in the message to be composed. Some other approaches addressing this issue are PAULINE [9] and the approach by Bateman and Paris [1]. Hence, the motivations underlying the texts produced are time constraints and other pragmatic parameters, and effective term selection techniques based on the addressee's role and command of domain knowledge. However, there is no evidence in these systems about the relation between the pieces of text actually generated and the precise communicative effect they might achieve ([1] constitutes an exception in this sense). As [14] have pointed out, the assumption underlying many of today's text planners³ that discourse is composed of hierarchically structured segments in a completely recursive way is inadequate for extended explanations. Hence, if saying 'X' is assumed to be comprehensible, and saying 'Y' as well, then it is concluded that saying 'X and Y' is also a suitable message. Within a certain range, this assumption is reasonable (for instance, mentioning the track in addition to a train's departure time is perfect), but this is clearly not without limitations (for instance, the description of a non-trivial route may easily become too complex).

Hence, generation systems never have developed a quantified model of the communicative effort involved in understanding their messages - but this has not been necessary so far: The situation description from which a communicative act is to be produced is hardly detailed and accurate enough to allow for the rich variety of reactions humans are able to produce under comparable circumstances, and the amount of information to convey is hardly ever so large that techniques to maintain the addressee's attention are a serious concern.

³ This assumption is based on Cohen and Levesque's model of rational action and interaction [3].

3. TOWARDS A FORMAL MODEL

3.1. Choosing Degrees of Explicitness

In general terms, the goal in selecting a suitable degree of explicitness and implicitness lies in following the concept identified in the course of the experiments carried out by [18], namely leaving "direct causes" to be inferred by the reader, and expressing "indirect causes" explicitly. This strategy tends to achieve a balance between pieces of information expressed explicitly and others left implicit, in the spirit of the discussion of explanations given in section 1.

In the approach adopted in [8], we have envisioned this goal by exploiting inference patterns based on the relations between generic regularities and individuals involved in principle-based explanations, thereby also taking scalar implicature [6] into account. This reasoning is based on the relevance that each of these pieces of information bears in an explanation context for understanding the rationale behind the decisions made (see also the examples in sections 1 and 3.2). As argued in [8], the mechanism developed is extendable to capture at least the following types of inferences:

- Inferring *plausible sequences of actions*; by exploiting expectations about intermediate steps in descriptions of action sequences can and should be omitted to produce less verbose and more natural text (see [12] and [18]).
- Inferring *causal relations between action and their results*; reference to a problem-solving step to be accomplished can be established by the action to be performed or by the state to be achieved, according to the linguistic repertoire available (see [17]).

In our application, the distinction between "direct" and "indirect causes" is implemented by leaving the results of chained inference steps implicit under particular circumstances only: the results of scalar implicature (and logical deduction) are left to be uncovered by the addressee's inference capability, even if combined with other inferences to be made. Eventually, this strategy may be refined if the effect of default expectations in the domain at hand is better understood and can be captured more systematically.

We conceive that it is also possible to pursue other strategies, either by producing more concise utterances and thereby increasing the burden on the addressee's inferential capability, or by expressing more inference steps explicitly. However, we do not go into potential variations of this aspect here, we simply adopt the approach of selecting the "best" balance between expressing information explicitly or implicitly, on the basis of the heuristics incorporated.

3.2. Deciding Upon Degrees of Detail

In order to decide whether the complexity of a message is beyond what the addressee can reasonably be assumed to understand with convenience, we propose the computation of three measures that serve as partial indications for this purpose: (1) The number of *entities* entailed in the propositional representation of the message; (2) The number of *predicates* used to build assertions about these entities; (3) The number of *relations* holding between entities and predicates.

In order to define entities and predicates for this purpose, we rely on the domain ontology of the underlying world model. The motivation behind lies in the assumption that, if the communicative purpose of the message is recognized, the conceptualization of the addressee should mirror to a large extent this categorization. In the office planning domain, for instance, 'employees', 'rooms', 'resources' constitute entities, while 'next-to', 'assigned-to' constitute predicates.

As a total measure for complexity we propose the formula $\max(e+p, r, 2*i)$, where p, e, r, i represent the number of entities, predicates, relations, and inferences, respectively. In the texts we are dealing with, these factors tend to be roughly the same; by the consideration of all factors, we intend to treat also other types of texts adequately: in enumerations, for instance, the number of relations decreases in comparison to the number of entities and predicates involved, while it increases if network-like relations between a set of entities and a set of predicates are to be expressed.

Returning back to the examples discussed in section 1, we are able to justify our assessments in some formal sense. Figure 2 illustrates the complexity measures computed for each of these sentences, confirming the intuitive assessment made in section 1: the measurements introduced indicate a fair correspondence with Miller's magical number seven.

The categorization into what constitutes an entity and what constitutes a predicate is problematical, since there are several sources of inaccuracies in this approach: for instance, complications due to vague expressions and due to quantification relations, and conceptualizations made by the addressee, which may differ significantly from those embodied in the system's conception consisting of entities and relations. We may encounter this problem by incrementing the calculations in case terminological transformations or lexical operations involve significant changes in the representation of the message. Eventually, the calculation schema must be augmented in case involved quantification relations occur more frequently than in our application.

In cases where the complexity assessments indicate potential comprehension problems on behalf of the addressee, several strategies are possible: a simple, but risky one, in which the complex utterance is fully elaborated, and two more sensible strategies. One of them consists in reducing the degree of detail in a coherent way (i.e., not leaving out details arbitrarily, but orienting this reduction on structural regularities), especially if demanded by time and/or space limitations, and the other one consists in reorganizing the text so that additional structuring hints in the text produced enable the addressee to perform intermediate conceptualization already with parts of the information presented.

sentence number	1)	2)	3)	4)	5)
number of entities (e)	4	6	10	7	6
number of predicates (p)	1	1	3	3	1
number of relations (r)	4	6	13	9	7
number of inferences (i)	0	8	3	1	1
total = max (e+p, r, 2*i)	4	16	13	10	7

Figure 2: Complexity assessments for some example sentences

The choice among alternatives should preferably be interest-based. For instance, the choice between alternatives 4) and 5) in section 1 may be made in favor of text 4) if the user wants to vary the relative significance of constraints in defining problem specifications (therefore, he/she wants to know which ones prove relevant for the aspect in question). Alternative 5) is preferable for a user who wants to learn the rationale behind system decisions, which particularly includes associating justifications with constraints.

4. CONCLUSION

A mechanism assessing the comprehension effort of the addressee of a text or an utterance certainly seems to be of interest for theoretical models of natural language generation processes. While the proposal entailed in our method can hardly be considered a psychologically motivated approach, we believe that it contributes to the understanding of what ingredients a performance-oriented model of natural language generation should consist of, and how interaction between these ingredients can suitably be organized.

Apart from a purely theoretical perspective, we believe that these considerations are also relevant in practice, even if an assessment of the comprehension effort of the addressee does not manifest itself in the majority of generator programs themselves - which also may not even prove necessary in a good deal of applications. Hence, for simpler, eventually application-oriented systems, taking these considerations into account will help in making explicit the assumptions underlying the simplifications embodied, so that conditions for an eventual transfer to other domains become more evident - hence, the assumptions must hold across domains. However, it will be important to know about a system's limitations, where they manifest themselves, and when they are likely to weaken the system's usefulness.

Conditions under which assessing the complexity of a message to be generated may prove beneficial, hold in at least the following types of application:

- Presenting a significant quantity of records selected from a database; when envisioning an ambitious, flexible presentation, summary facilities play a crucial role based on suitable assessments in the spirit of our method.
- Generating business reports including a considerable amount of (heterogeneous) data, where summaries also play an important role; in that genre, it is the contextual inferability of information from some key propositions conveyed, which constitute the main difficulty. Hence, a quality assessment function has to take this particular aspect into account, whereas other influences on the ease of comprehension can be widely neglected.
- Explaining an issue of at least moderate complexity (for instance, presenting a chain of rules or a set of constraints with underlying justifications) in explanation generation; all aspects neglected in report generation (degree of detail, structure, acquaintance with terms) bear a significant amount of relevance here, and they should be incorporated in a suitable quality function.

REFERENCES

- [1] J. Bateman, C. Paris: *Phrasing a Text in Terms the User can Understand*. In Proc. of IJCAI-89, pp. 1511-1517, Detroit, 1989.
- [2] A. Cawsey: *Generating Explanatory Discourse*. In Current Issues in Natural Language Generation, R. Dale, C. Mellish, M. Zock (eds.), pp. 75-102, Academic Press, 1990.
- [3] P. Cohen, H. Levesque: *Speech Acts and Rationality*. In Proc. of ACL-85, pp. 49-60, 1985.
- [4] R. Dale: *Cooking Up Referring Expressions*. In Proc. of ACL-89, pp. 68-75, Vancouver, 1989.
- [5] A. Garnham: *Testing Psychological Theories about Inference Making*. In Memory and Cognition 10(4), pp. 341-349, 1982.
- [6] J. Hirschberg: *A Theory of Scalar Implicature*. Garland Press, New York, 1991.
- [7] H. Horacek: *Exploiting Conversational Implicature for Generating Concise Explanations*. In Proc. of EACL-91, Vol. 1, pp. 191-193, Berlin, 1991.
- [8] H. Horacek: *How to Avoid Explaining Obvious Things (Without Omitting Central Information)*, to appear in ECAI-94.
- [9] E. Hovy: *Pragmatics and Natural Language Generation*. In Artificial Intelligence, 43, pp. 153-197, 1990.
- [10] T. Jameson, W. Wahlster: *User Modeling in Anaphora Generation: Ellipsis and Definite Description*. In Proc. of ECAI-82, pp. 222-227, Orsay, 1982.
- [11] M. Just, P. Carpenter: *A Capacity Theory of Comprehension: Individual Differences in Working Memory*. In Psychological Review 99, pp. 122-149, 1992.
- [12] W. Kintsch, J. Keenan, G. McKoon: *Memory for Information Inferred During Reading*. In The Representation of Meaning in Memory, W. Kintsch (ed.), Earlbaum, Hillsdale, New Jersey, 1974.
- [13] G. Miller: *The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information*. In Psychological Review 63, pp. 81-97, 1956.
- [14] D. Mooney, S. Carberry, K. McCoy: *The Generation of High-Level Structure for Extended Explanations*. In Proc. of COLING-90, pp. 276-281, Helsinki, 1990.
- [15] J. Moore, W. Swartout: *A Reactive Approach to Explanation*. In Proc. of IJCAI-89, pp. 1504-1510, Detroit, 1989.
- [16] E. Reiter: *Generating Descriptions that Exploit a User's Domain Knowledge*. In Current Issues in Natural Language Generation, R. Dale, C. Mellish, M. Zock (eds.), pp. 257-285, Academic Press, New York, 1990.
- [17] D. Rösner: *Intentions, Rhetoric, or Discourse Relations? A Case from Multilingual Document Generation*. In Intentionality and Structure in Discourse Relations, Workshop held at the ACL-93, O. Rambow (ed.), pp. 106-109, Columbus, Ohio, 1993.
- [18] M. Thüning, K. Wender: *Über kausale Inferenzen beim Lesen*. In Sprache und Kognition 2, pp. 76-86, 1985.
- [19] J. Weiner: *BLAH: A System which Explains its Reasoning*. In Artificial Intelligence, 15(1), pp. 19-48, 1980.