A Discourse Model for Gist Preservation

Lucia Helena Machado Rino
DC - UFSCar
Caixa Postal 676
13565-905 São Carlos - SP, Brazil
Fax: (016) 2748233
lucia@dc.ufscar.br
Donia R. Scott
ITRI - University of Brighton
Lewes Road
Brighton BN2 4AT, UK
Fax: (00-44-1273) 642908
Donia.Scott@itri.bton.ac.uk

Abstract. This paper describes an approach to gist preservation during automatic summarization whereby the source is a complex information structure which must be "pruned" and organized in such a way as to make it appropriate for textual expression. Based on a discourse model, we propose a process whereby gist is guaranteed at the deep level according to communicative and rhetorical settings. The main function of such a goal-driven summarization model is to map intentions onto coherence relations whilst still observing the semantic dependency indicated by the message source. The discourse model is thus based on an association of intentionality, coherence and semantics, which guides the production of summary message sources that highlight the central proposition of the discourse.

Keywords: Automatic summarization, discourse modeling, text generation.

Introduction

This paper addresses the processes underlying automatic summarization within the context of natural language generation. In particular, the main problem under focus is how to derive the text plan of a summary of a complex information structure, i.e., a reduced but suitably enhanced structure that eliminates redundant or less relevant information while preserving the gist of its source. This problem can be viewed within the context of a traditional natural language generation system to which we add a summarization component. Since the focus of this work is in the underlying processes, a summarizer is simply seen here as the text structurer. In this special context, the input (i.e., the complex information structure) is referred to as *primary message source*, and the output is a new, concise, text plan derived from a subset of such a structure and overlaid with a discourse structure, henceforth named *summary message source*. This source can then be passed to a realization component for textual expression.

In this paper, we will not address the later stage of linguistic realization. Instead, we will discuss two main issues concerning the production of summaries in the specified context:

- the required characteristics of the primary message source for it to be amenable to the derivation of coherent and accurate summaries that are gist-preserving;
- the nature of the process required to derive summary message sources.

Main concepts

A discourse is typically viewed within the computational linguistics community as a hierarchical composition of discourse segments represented linguistically by a structured collection of clauses [Grosz and Sidner, 1986, Hobbs, 1985, Hovy, 1993]. Discourse segments are seen as units of meaning (e.g., sentences or propositions), each having a specific function in the discourse, established by the inter-relationship among its elements. In general, each discourse segment reflects a subtopic of the global discourse topic and its relation to the overall communicative goal by means of its own communicative function in the discourse.

At the deep level of processing, the main elements manipulated or produced by our summarizer are the following:

The communicative goal. This is the primary intention of the writer in producing the text.

The central proposition. This is the "kernel" of the discourse [Levelt, 1989], i.e., the information around which the discourse is organized to satisfy the communicative goal. This information can be highly complex. For the purpose of this paper, however, we restrict it to a single component of the primary message source, i.e., a leaf of its information structure, as it will be shown shortly.

The primary message source. This is a complex information structure composed of three distinct kinds of information: a communicative goal, a central proposition, and a knowledge base. This base is a hierarchical structure of semantically inter-related content units, as illustrated in Fig. 1 (underlined words stand for the super-components of the knowledge base), and is the only component of the primary message source that is available for condensation. Such a structure can be "pruned" by extracting those propositions that are superfluous in the specific context determined by the communicative goal and the central proposition. Fig. 1 shows the structure of the knowledge base for the following example text (numbered clauses are propositions, but not all of them have been exhaustively dismembered. Occurrences of, e.g., X, Y or Z in the message source stand for implicit concepts whose content could not be retrieved from the text surface):

Example text:

1. This paper presents the results of measurements on the keyhole structure produced during the laser processing of materials.

2. A cw CO_2 laser was used to produce keyhole structures in water and measurements were made of the depth of the keyhole as a function of laser power and the force due to gravity.
3. The data show that as the laser power is increased, the keyhole depth also increases.
4. However, changes in the force due to gravity seem to have little effect on the depth.
5a. The data for the different power conditions are compared with calculations using the Andrews and Atthey expression for the hole shape
5b. and good agreement is achieved.
6. There is however no term in the theoretical description which accounts for the downward flow of liquid around the keyhole.
7a. This flow is estimated at 20 cm/s
7b. and is thought to be an important factor
7c. which it is planned to include in more detailed calculations.
8. Further measurements are also planned.

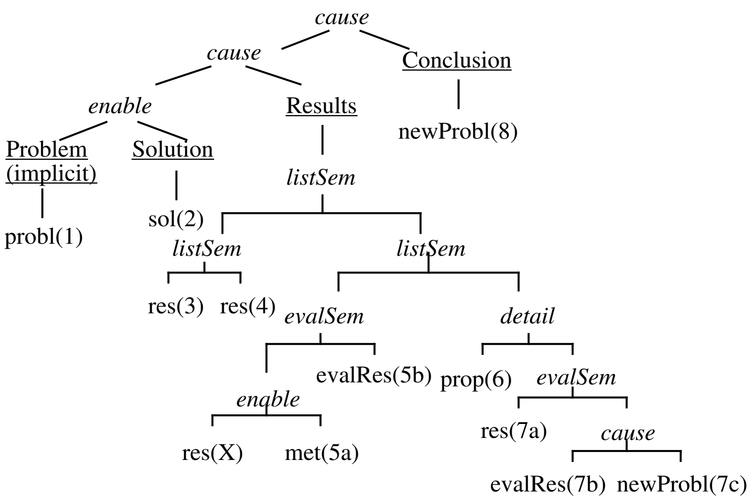

Fig. 1. A knowledge base of a primary message source

The summary message source. This is a discourse structure which reflects the logical chaining of ideas and propositions built around the central proposition by means of subthemes, or subtheses. It is derived from the primary source by selecting only the relevant material contained within it and organizing that material coherently. Selection and organization are pipelined, so that first the knowledge base is pruned and then the discourse of the summary is organized according to the resulting knowledge base, the communicative goal and the central proposition. Fig. 2 and Fig. 3 show, respectively, the condensed knowledge base and the summary message source to describe a result in a summarizing process whereby the primary message source is composed of the communicative goal Describe and the central proposition *result "7a"* (or *res(7a)*, as shown). Bold branches in the summary message source indicate nuclear subtrees, according to the *Rhetorical Structure Theory* (RST) [Mann and Thompson, 1987]. In the model presented here, the propositions of the condensed knowledge base will be used to complement or emphasize the central proposition, at the same time that they contribute to the communicative goal [Grosz and Sidner, 1986, Scott and Souza, 1990]. According to the above, the central proposition of a discourse will be one of the leaves of both, the primary and the summary message sources. If the summary is gist-preserving, the central proposition of both sources will be identical. A possible hand-generated summary based upon the illustrated summary message source is given below.

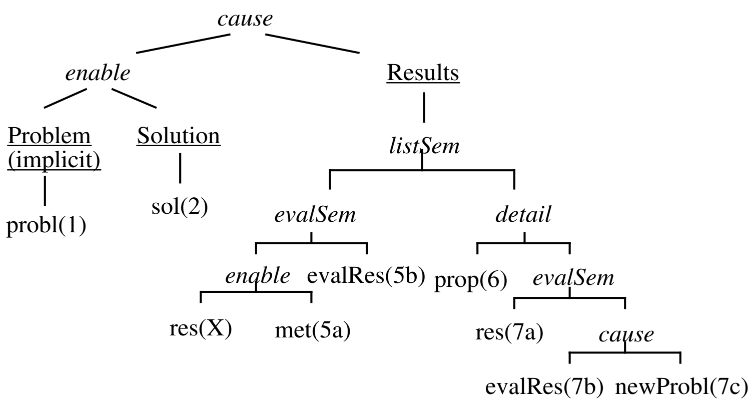

Fig. 2. A condensed knowledge base

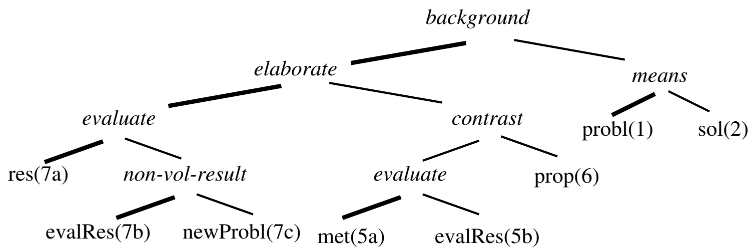

Fig. 3. A summary message source of a description

Example summary: A possible description of a result

The downward flow of liquid around a keyhole structure produced during the laser processing of materials has been estimated at 20 cm/s, an important factor which must be included in more detailed calculations. Although good agreement is achieved with a comparison between data for the different power conditions and calculations using the Andrews and Atthey expression for the hole shape, there is no term in the theoretical description which accounts for such a flow.

Strategic account

The results described in this paper are based on the analysis of a corpus of scientific journal articles from a single technical domain (physics). Given the highly stereotyped nature of this genre, our analysis revealed a well-known pattern of discourse organization, known as Problem-Solution (P-S) [Hoey, 1983]. This pattern reveals the set of common semantic relations used in the primary message source [Rino and Scott, 1994]. Many of them can be fairly straightforwardly mapped onto coherence relations drawn from existing theories of discourse organization, e.g., RST and Hobbs' coherence relations. The analysis also identified the most frequent communicative goals (Describe, Relate, and Discuss). For an intentional account, we add to our model the *Grosz and Sidner Discourse Theory* (GSDT) [Grosz and Sidner, 1986], whose dominance and precedence relations are independent of a linguistic behavior and characterize the search for a communicative setting at a high level of abstraction.

GSDT and RST are complementary and both allow for a functional approach. RST specifies the intended effects of a coherence relation on the behavior of the reader and expresses (albeit coarsely) the underlying discourse intentions, conveying the propositional content in a relational form that is essential for coherence. GSDT, on the other hand, focuses fundamentally on the intentional contribution of individual discourse segments to the overall discourse by allowing for structural relations between intentions. We show here that this theory can lead to the choice of specific coherence relations, complementing the lack of a clear specification of communicative purposes of the RST. At the same time, chosen coherence relations cannot violate the conceptual dependence specified in the primary message source. The inter-relationship between intentions and coherence relations is critical to summarization: it is this that provides the basis for determining the partial contribution of content information to gist preservation and the satisfaction of the overall communicative goal.

Modeling discourse

Formal and functional perspectives both hold in our discourse modeling: we make use of a formalism that summarizes the primary message source by considering the functional perspective, responsible for the discourse internal structure and for the relationship between the intentional and coherence levels. Therefore, discourse functionality determines discourse organization by mapping intentionality

Semantic relations	Coherence relations	Structural intentions among intentions	Cases
enable(Y,X)	purpose1(X,Y) means(Y,X)	X sat-precedes Y Y dom X	1
rationale(X,Y)	purpose2(X,Y) justify1(X,Y)	Y supports X X dom Y not X sat-precedes Y	2
proof(X,Y)	evidence(X,Y) justify2(X,Y)	Y sat-precedes X X dom Y Y supports X	3
cause(X,Y)	non-vol-result(Y,X) non-vol-cause(X,Y)	Y sat-precedes X X dom Y Y generates X	4
listSem(X,Y)	list(X,Y) contrast(X,Y)	X symmetr Y	5
attribute(X,Y) *detail*(X,Y) *exemplify*(X,Y)	elaborate(X,Y)	Y supports X	6
evaluateSem(X,Y)	evaluate(X,Y)	X dom Y	7
reason(X,Y)	explain(X,Y)	Y generates X	8
sequence(X,Y)	sequence(X,Y)	X sat-precedes Y	9
background(X,Y)	background(X,Y)	not X sat-precedes Y Y supports X	10

Table 1. Mapping of intentions and semantic relations onto coherence relations

onto coherence, while still observing the semantics of the knowledge base to provide the informative means of the discourse.

Modeled in such a way, discourse is organized considering three distinct levels of structural relations: the functional one, which combines intentions and coherence relations, the informative one, given by the knowledge base, and the formal one, which addresses linguistic resources based upon the interaction between informativity and functionality. This formal method of producing discourse is illustrated in Table ??. For example, in Case 5 there is a formal correspondence that reads: *If a "symmetry" intentional relation holds between discourse segments X and Y, and if both segments are in a "listSem" (i.e., a parallel) relation in the knowledge base, then the preferrable coherence relations between X and Y are either* LIST *or* CONTRAST.

Table ?? addresses all the possible mappings of our current discourse model, i.e., all three sets of relations are necessary and sufficient for summary planning. Semantic relations are naturally depicted in the English language. Coherence relations are determined through the propositional delimitation of spans of sample texts. Structural relations between intentions follow the GSDT. In addition to these, we find it necessary to include the relation symmetry, which holds between two intentions when one contributes to the other and vice-versa. Of the coherence relations relevant to our task, only two present a symmetric

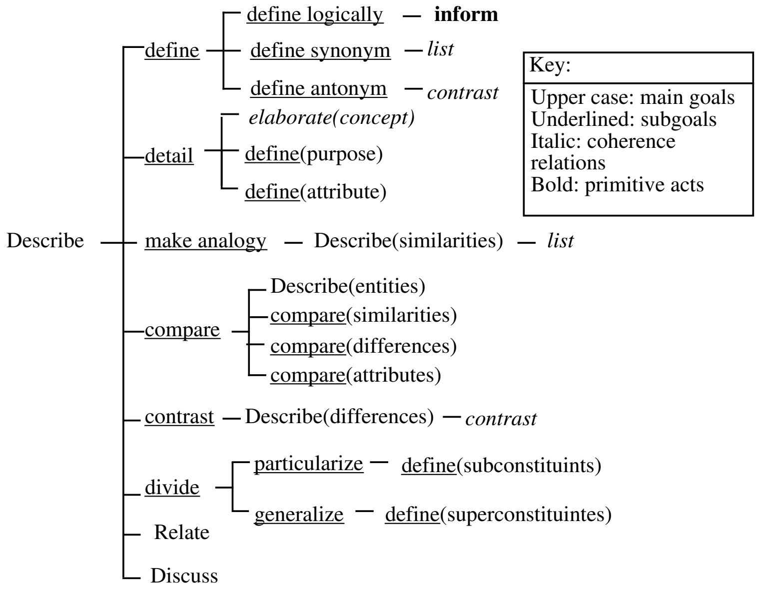

Fig. 4. Communicative goals and their interaction with coherence relations

account at the intentional level: LIST [Scott and Souza, 1990] and CONTRAST [Mann and Thompson, 1987], as shown in Case 5.

Along with the main communicative goals, there are other subgoals whose interplay display a complex interaction among different text types. For example, narratives exchange features with expositions when they narrate facts of an explanation of plans. Such a complex interplay can provide a *systemic network*, which allows for the mapping between communicative goals and coherence relations in a special way, as partly shown in Fig. 4.

Conceptual classes found in the corpus analysis determine another kind of network, i.e., a *communicative inheritance network*, which details further the communicative goals. For example, an entity description can be a description of a solution or a problem, which are two conceptual classes of the scientific discourse model. Such classes are used to annotate the content units of a knowledge base, e.g., "Result" produces the annotation "res" in Fig. 1.

Computational account

Both inheritance and systemic networks allow for the definition of discourse schemata [McKeown, 1985] and plan operators [Moore and Paris, 1993]. A schema guarantees that the proper communicative setting is initially established, in order to preserve gist. For example, given that a super-structure of a summary can be represented by the logical sequence *Situation-Problem-Solution-Results-Discussion and evaluation of results* [Hoey, 1983], to describe a specific result a schema is chosen which emphasizes either the subpath *Describe-Feature,Describe-Attribute,Describe-ValueOfResult* or the subpath *Describe-Object,Describe-Result*. Plan operators allow for distinct rhetorical settings based upon different choices of rhetorical strategies (see Fig. 4).

Gist preservation is addressed at all levels of discourse processing. The most important resources to address this issue are the plan operators, which are responsible for discourse organization, in that they map intentional and semantic relations onto coherence relations, according to the correspondence shown in Table ??. Graduallly, the summary message source is constrained to present the central proposition as the "most nuclear" unit.

We follow Moore and Paris to specify our plan language. Communicative goals can be represented, for example, by *Describe, Interpret, or Discuss*, and can be associated to the writer's intention to make the reader believe, or act according to, a proposition. A discourse is successfully produced when it achieves the intended change of the mental state of the reader. Linguistic goals to support these intentions can be primitive acts (only inform in our current model) or rhetorical goals (e.g., evidence, or justify), which are directly expressed as coherence relations.

Plan operators handle both types of goals. However, only those that explicitly post linguistic goals can modify the discourse structure under construction. Planning is complete when all the active plan operators are successfully applied, generating all the leaves of the discourse structure, i.e., the primitive acts given by *inform* nodes. At this point, the proof space resulting from the recursive and progressive account of constrained discourse organization registers not only the thread of discourse, but also the space of contribution of each discourse segment to the overall discourse, in terms of the purpose of each particular segment.

Each plan operator is a modified RST plan (embedding functional aspects of discourse organization) whose satellites can be essential or optional. Essential satellites affect the purpose of the nucleus and, consequently, the contribution of the corresponding discourse segment to the overall communicative purpose. For this reason, they must be further developed in the plan. Optional ones, on the other hand, may elucidate the nucleus, but are not crucial. The decision to include them in a summary plan depends on the constraints (for example, the degree of detail required). Such satellites are responsible for varying degrees of conciseness at the deep level of summarization.

Since subgoals can be either communicative or linguistic, coherence relations are determined only when a plan operator applies to a context that provides appropriate conditions for the mapping between intentions and semantic relations. This approach proves appropriate for summarization in that it provides the flexibility to choose plan operators that can convey and organize information in varied ways, provided that the central proposition is preserved and is guaranteed to be available to the realization phase. Moreover, built on a constraint satisfaction model, the final discourse structure is guaranteed to achieve the communicative goal because it operates on partial information, gradually constraining the range of possible choices of new information to compose such a structure.

Discussion and Conclusion

Other work has explored a similar communicative account to the one presented here, e.g., in text generation by Cawsey (1990) and Maybury (1992), and in automatic summarization by Sparck Jones (1993). However, none have addressed the crucial problem of preserving gist in the context of condensing the message to be conveyed. Pursuing this view, we add to Moore and Paris' discourse model the specific constraints for gist preservation and prominence, assigning the position of the central proposition of a summary message source to the "most nuclear" leaf of an RST-like discourse structure. In addition, we trigger the planning process by narrowing the possible paths to achieve the communicative goal based on the typical semantic organization of the primary message source. The consistency provided by the highly stereotyped discourse under investigation allows us to identify the possible super-components that involve the central proposition of the discourse to be constructed. In this respect, the Problem-Solution pattern is particularly interesting, allowing for schemata definitions that indicate the possible ways of satisfying the summarization constraints.

The correspondence between semantics and coherence described here is based on an empirical study. From Table ??, we can see that the mapping of semantic relations to coherence relations can be one-to-many or many-to-one, and that coherence relations do not necessarily map to the same structural relations at the intentional level (e.g., PURPOSE and JUSTIFY). This is fortuitously advantageous for summarization, in that it provides the necessary flexibility to produce different discourse structures for the same message and also varying degrees of conciseness or informativeness.

Compared to other approaches based on schemata (e.g., [McKeown, 1985]), the approach described here provides more flexibility through the addition of plan operators which combine in a single formalism the interdependence between three distinct levels of discourse representation, namely intentionality, coherence and semantics. None of these structural levels taken alone is able to address any of the others [Sparck Jones, 1993], e.g., a communicative setting does not completely account for a rhetorical one; semantic relations do not express intentions by themselves. Besides such advantages, "reasoning" provided by plan operators is not predetermined: special cases where the choice of a specific discourse relation runs the risk of obscuring the central proposition can be handled by the constraints field of any plan operator.

Our approach relies on the active identification of information that can complement the central proposition to be conveyed. However, it takes an inverse view to that proposed by Robin (1994). Whereas we identify less relevant propositions, *omitting* them from the primary message source and reorganizing them, Robin identifies relevant propositions, *adding* them to a given proposition in order to build up a text. In our approach, automatic summarization is based on conventional text generation processes for modeling discourse, to which we add a severe constraint: the preservation of the gist of a message source. We show that a communicative account is essential to determine the space of contribution of discourse segments to both gist preservation and the overall communicative goal. In this, our work supports the insights provided by Sparck Jones.

In carrying out this work, we have provided a specification of the correspondence between intentions and coherence relations on the basis of the highly stereotyped discourse of science. We have applied our discourse model to a small number of scientific domains and to another genre (technical texts) [Rino, 1996]. Our results so far show that gist can be preserved in summary generation provided that the corresponding primary message sources are structured according to the Problem-Solution pattern. This pattern is not specific to scientific texts, but can be found in a number of other genres, e.g., [Jordan, 1992], suggesting that our discourse model for gist preservation may be more generally applicable to a wider context of natural language generation.

Acknowledgments. The research described in this paper was partially supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Project No. 201610/92-2.

References

[Cawsey, 1990] Cawsey, A. Generating Explanatory Discourse. In R.Dale; C. Mellish and M. Zock (eds.), *Current Research in Natural Language Generation*, pp. 75-101. London, Academic Press, 1990.

[Grosz and Sidner, 1986] Grosz, B. and Sidner, C. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, Vol. 12, No. 3, 1986.

[Hobbs, 1985] Hobbs, J.R. *On the Coherence and Structure of Discourse.* Technical Report CSLI-85-37, Center for the Study of Language and Information, Stanford University, 1985.

[Hoey, 1983] Hoey, M. *On the Surface of Discourse.* George Allen & Unwin (Publishers) Ltd., 1983.

[Hovy, 1993] Hovy, E. Automated Discourse Generation Using Discourse Structure Relations. *Artificial Intelligence* 63, pp. 341-385, 1993.

[Jordan, 1992] Jordan, M.P. An Integrated Three-pronged Analysis. In W.C. Mann and S.A. Thompson (eds.), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* John Benjamins Publishing Co., 1992.

[Levelt, 1989] Levelt, W.J.M. *Speaking: From Intention to Articulation.* The MIT Press, Cambridge. 1989.

[Mann and Thompson, 1987] Mann, W.C. and Thompson, S.A. *Rhetorical Structure Theory: A Theory of Text Organization.* Technical Report ISI/RS-87-190, June 1987.

[Maybury, 1992] Maybury, M.T. Communicative Acts for Explanation Generation. *Int. Journal of Man-Machine Studies* 37, pp. 135-172, 1992.

[McKeown, 1985] McKeown, K.R. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text.* Cambridge University Press, 1985.

[Moore and Paris, 1993] Moore, J.D. and Paris, C. Planning Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information. *Computational Linguistics*, Vol.19, No. 4, pp. 651-694, 1993.

[Robin, 1994] Robin, J. *Revision-Based Generation of Natural Language Summaries Providing Historical Background: Corpus-Based Analysis, Design, Implementation and Evaluation.* PhD. Thesis. Columbia University, December 1994.

[Rino, 1996] Rino, L.H.M. *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos.* PhD. Thesis. Universidade de São Paulo, Brasil. Abril 1996.

[Rino and Scott, 1994] Rino, L.H.M. and Scott, D. Content selection in summary generation. *XI Simpósio Brasileiro de Inteligência Artificial.* Fortaleza, Brazil, October 1994.

[Scott and Souza, 1990] Scott, D.R. and Souza, C.S. Getting the Message Across in RST-Based Text Generation. In R.Dale; C.Mellish and M. Zock (eds.), *Current Research in Natural Language Generation.* London, Academic Press, pp. 47-73, 1990.

[Sparck Jones, 1993] Sparck Jones, K. *Discourse Modelling for Automatic Summarising.* Tech. Rep. No. 290. University of Cambridge, Feb 1993.

This article was processed using the LaTeX macro package with LLNCS style