DMSumm: Review and Assessment*

Thiago Alexandre Salgueiro Pardo and Lucia Helena Machado Rino
{taspardo,lucia}@dc.ufscar.br
Departamento de Computação, Centro de Ciências Exatas e de Tecnologia
Universidade Federal de São Carlos
Rodovia Washington Luiz, km 235 – Monjolinho
Caixa Postal 676
13565-905 São Carlos - SP

Abstract. In this paper we review DMSumm, an automatic summary generator based on a discourse model that combines semantic, rhetorical and intentional knowledge. We assess the automatic results in the light of three basic constraints: gist preservation, communicative goal satisfaction and textuality.

1 Introduction

In this paper, we pinpoint some of the main aspects of the DMSumm system, the Discourse Modeling SUMMarizer described in (Pardo and Rino, 2001), predominantly focusing on discourse organization. This addresses a special three-level text planner (Rino, 1996a), bringing together intentional (Grosz and Sidner, 1986), rhetorical (Mann and Thompson, 1987; Hobbs, 1985) and semantic relations. These, in turn, address information content on the basis of the Problem-Solution (P-S) model (Winter, 1977; Jordan, 1980) and resemble Jordan's (1992) clausal relations. We thoroughly show how discourse production is carried out (Section 2), illustrating the DMSumm reasoning (Section 3). Then, we assess the automatically produced summaries (Section 4), in order to discuss the DMSumm potentialities (Section 5).

2 Summary Generation

Summary generation, here, refers solely to the problem of taking an *input message* and recognizing its most relevant information to appear in a summary, organizing it according to communicative goals, and realizing the resulting summary plan into the text itself. This addresses a 3-step pipelined text generator, whose input message is already an internal representation of a source text, as it is shown by the DMSumm architecture in Fig. 1. The input message is a composition of three information units: two single ones – the Central Proposition (CP) and the Communicative Goal (CG) – and a complex, semantically structured one – the so-called Knowledge Base, or KB. In limiting DMSumm to this setting, interpretation is assumed to have been carried out previously, and, actually, it has been so far carried out by hand, by human specialists. Defined in this way, the input indicates that the main components of discourse production are *what brings about the primary discourse motivation* (the CP), *what depicts the*

* This work has been funded by FAPESP – Fundação de Ampara à Pesquisa do Estado de São Paulo.

intertwining of discourse segments, aiming at building up the discourse (the CG), and *what allows content to be handled* (the KB). Summary generation proceeds thus under three basic constraints, namely, gist, or CP, preservation (Constraint 1), CG satisfaction (Constraint 2), and textuality warranty[1] (Constraint 3) (Rino, 1996a). In Rino's model, CP is considered to be a single information unit. Conversely, gist is a more general concept, in that it involves not only the CP, but also CG. However, in our automatic summarization (AS) scenario, CP preservation has barely been dealt with on its own, for discourse production is also based upon CG satisfaction. Here, we will see how the referred constraints are handled in DMSumm, by briefly describing its processes, which are introduced in (Pardo and Rino, 2001). Readers should refer to that article for further details on the DMSumm description and its operation.

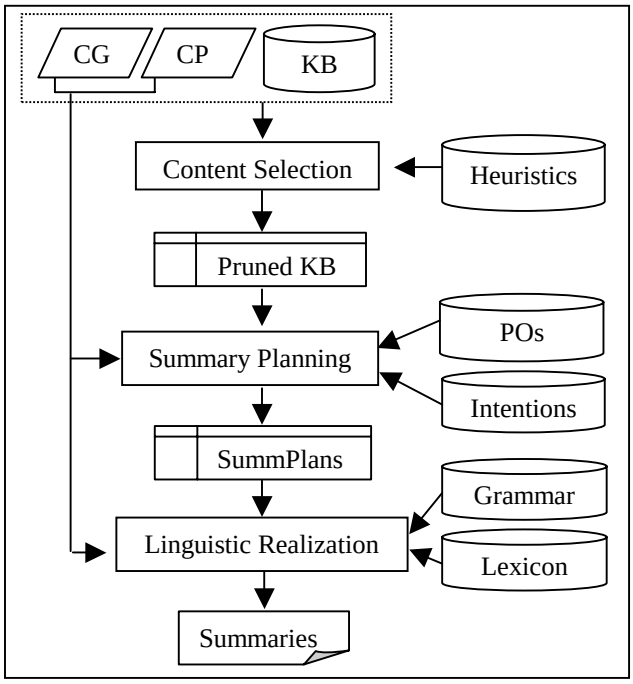


Figure 1: DMSumm Architecture

2.1. Content Selection

This heuristic-based process has been originally proposed by Rino and Scott (1994) and has been refined in DMSumm. The heuristics are based on the semantic structure of the KB, which takes after the P-S paradigm (Winter, 1977; Jordan, 1980). In this way, content selection addresses the P-S well-defined logical sequence of super-components. Additionally, it focuses upon information units that are semantically related through clausal relations (Jordan, 1992). The main underlying assumptions of such heuristics are the following: a) source texts usually address technical issues; b) readers are knowledgeable enough to handle implicit information; c) the P-S paradigm is reasonably common and can be found in a wide variety of texts, no

[1] Textuality, here, is understood as the property of a text being meaningful to the reader. For this reason, it addresses both, the coherence and cohesion of a text.

matter their genre or domain (Hoey, 1983; Rino, 1996b; Jordan, 2000); d) by delineating the KB information segments on either the P-S or the clausal account, the heuristics can address KB pruning by considering either of them; and, finally, e) the CP can vary according to the writer's intentions in focalizing information. Assumptions (d) and (e) are of utmost importance in DMSumm, for they provide the means to vary discourse structuring, even when the input message remains invariable. The heuristics are shown bellow, with justifications in brackets:

- H1: select only the *problem* and *solution* segments of a KB (provided that gist addresses either problem or solution, these segments are enough in a summary);
- H2: select every information but *results* (these should not convey substantial information);
- H3: exclude actions pre-conditions (readers can infer them);
- H4: exclude instruments and events purposes (readers can infer them);
- H5: exclude any proof segment of the KB (one should avoid too much detail);
- H6: exclude effect from causal relations (readers can infer them);
- H7: exclude objects attributes, details, or examples (these are not essential);
- H8: exclude evaluations (in technical and objective communication, personal opinions must be avoided);
- H9: exclude events reasons (in certain contexts, these may be superfluous);
- H10: exclude *situation* (avoid background information if shared with readers).

Pruning verifies that a) CP, which is a leaf in KB, be also a leaf in the resulting pruned KB (cf. Constraint 1); b) CG lead to intentions that delineate contributing discourse segments (cf. Constraint 2) and c) end up relating CP to other segments (cf. Constraints 1 & 3).

2.2. Summary Planning

This process builds upon the mapping of semantic and intentional relations onto rhetorical ones. Table 1 shows some examples of the constraints to be observed at both, semantic and intentional levels, to rhetorically organize a summary, considering that 1) semantic organization is provided as input in the KB; 2) the intentional account, i.e., the GSDT (Grosz and Sidner Discourse Theory) relations satisfaction-precedes (*sp*), dominates (*dom*), supports (*sup*) and generates (*gen*), and the Rino's relation symmetry (*symm*), is previously defined between every two KB potential information units. In DMSumm, this mapping is carried out by plan operators, or POs (Maybury, 1992; Moore and Paris, 1993), triggered by CGs. The only CGs addressed by Rino (1996a) are *describe*, *report*, and *discuss*, but they are further refined to observe the contributing GSDT discourse setting. In this way, diverse planning strategies have been devised.

Table 1: Mapping of intentions and semantics onto rhetoric

Semantics	Rhetoric	Intentions
enable(Y,X)	means(X,Y)	X sp Y, Y dom X
cause(X,Y)	nonvolresult(Y,X)	Y sp X, X dom Y, Y gen X

The steps below depict an interpretation of Table 1 in planning a summary, considering that we want to generate a MEANS relation between two discourse segments X and Y (X is the goal to be achieved through Y). Fig. 2 shows a PO that operates on such a mapping to report a concept X. Currently, DMSumm amounts to ca. 89 POs similar to that.

1. if there is an *enable* relation between information segments X and Y in the KB and
2. if *sp* and *dom* relations hold between them at the intentional level, then

3. generate a MEANS relation at the rhetorical level.

Name	report-by-MEANS
Header	report(X)
Effect	know(R,X)
Constraints	not know(R,X), enable(X,Y), sp(X,Y), dom(Y,X)
Nucleus	report(X)
Satellite	know(R,MEANS(X,Y))

Figure 2: Example of PO

2.3. Linguistic Realization

Linguistic realization has been so far simply undertaken by considering a template-based approach. This provides just canned text spans in the final summaries, linked by discourse markers. In linearizing a summary plan (SummPlan), a template choice takes place according to the focused rhetorical relation and this indicates the appropriate discourse markers (see Fig. 3, for the MEANS rhetorical relation). The rhetorically inter-related propositions are, thus, mapped onto the canned text spans and connected at the surface. This process is based on Marcu's work (1997), in that it specifies ordering and clustering constraints to linearize a summary plan. Ordering features pinpoint which text span comes before the other; clustering ones indicate whether text segments will involve a single sentence, adjacent sentences or even different paragraphs. In defining DMSumm templates, canned text spans are entirely and literally extracted from source texts, except when they involve context dependencies. These have been so far hand-resolved, for both clarity and keeping text spans independent from each other. This is particularly convenient for the template-based approach, but should be improved in a more elaborated linearizer.

To assess DMSumm, we have specified two linguistic realization engines by defining two template sets: one for English and another for Brazilian Portuguese (hereafter, referred to just by Portuguese). This process has not been troublesome, since we could reuse most of the English-based counterpart.

MEANS	
Ordering	nucleus before satellite
Clustering	single sentence
Discourse markers	*by means of*

Figure 3: MEANS template

3. DMSumm at Work

In this section, we explore automatic summarizing through the '*Using Computers in Manufacturing*' text (Jordan, 1980, p. 225). Its segmented version is shown below, considering that each text segment corresponds to an information unit, which in turn will be corresponding to a proposition at the discourse level. This segmentation strategy has been formerly undertaken by many other RST human analyzers (e.g., Marcu, 1997).

Using Computers in Manufacturing

1. Whether you regard computers as a blessing or a curse, the fact is that we are all becoming more and more affected by them.

2. Yet, in spite of this, the general level of understanding of the power and weaknesses of computers among manufacturing managers is dangerously low.

3a. In order to counteract this lack of knowledge, the Manufacturing Management Activity Group of the IprodE is organizing a two-day seminar on "Computers and manufacturing management"

3b. to be held at the Birmingham Metropole Hotel at the National Exhibition Centre from 21-22 March 1979.

4. The seminar has been specially designed by the IprodE for managers concerned with manufacturing processes and not for computer experts.

5. The idea is that delegates will be able to share the experiences of other computer users and lear of their successes and failures.

6. The seminar will consist of plenary sessions followed by syndicates where delegates will be arranged into small discussion groups.

Fig. 4 illustrates a KB corresponding to such a source text. Nodes in italics represent semantic relations; underlined ones signal P-S super-components, and leaves indicate the basic information units, already annotated by P-S tags. It is important to stress that, although such a structure resembles a rhetorical one, it is purely informative, for it relates content on a domain basis, and not on a discourse basis.

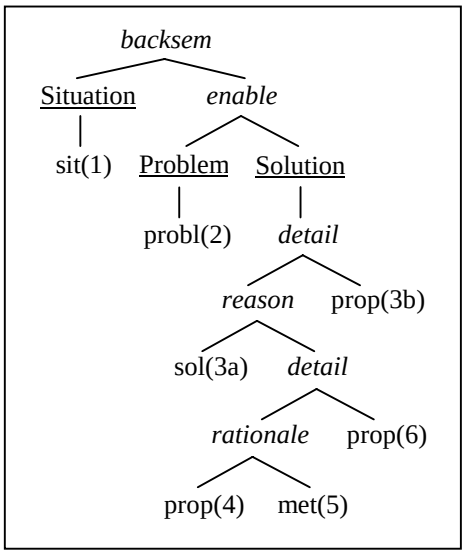


Figure 4: KB

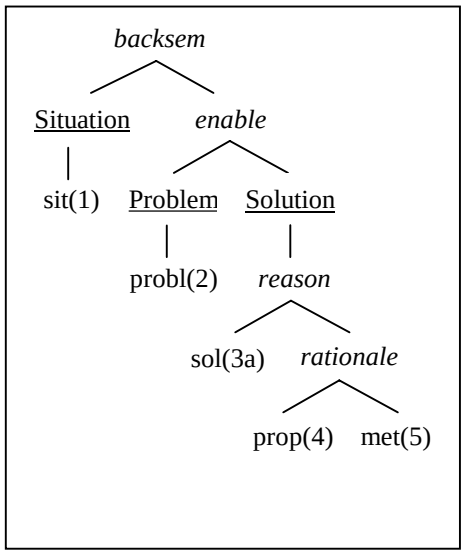


Figure 5: Pruned KB

To illustrate how DMSumm works, we assume that we want to report a problem. So, the DMSumm input message is completely defined by the following three components: the KB, the CP = probl(2), and the CG = report. Fig. 5 shows the pruned KB when we apply, e.g., heuristic H7 during content selection. Along with CP and CG, such a KB amounts to the condensed input message to the summary planner, comprising the very same input CP and CG, but now the pruned KB (refer to Fig. 1). By applying every possible DMSumm planning strategy, we generated 11 summary plans for this example. Two of them are shown in Figs. 6

and 7, along with possible linguistic realizations. They make evident that different degrees of compression are feasible, always preserving the CP at the leftmost nuclei.

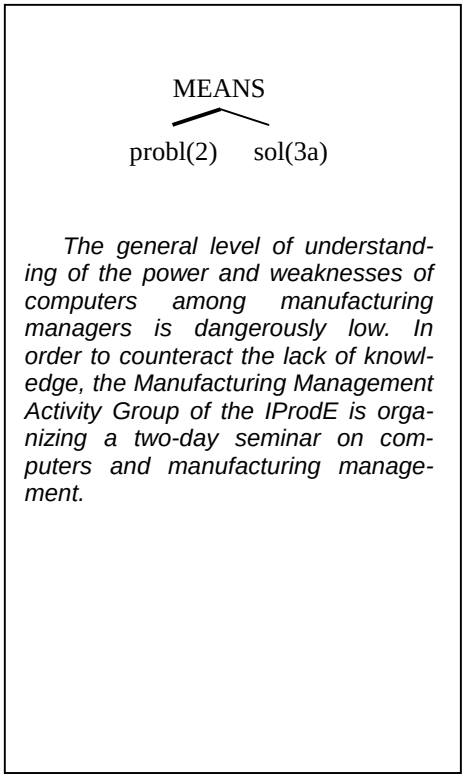


Figure 6: Plan 1

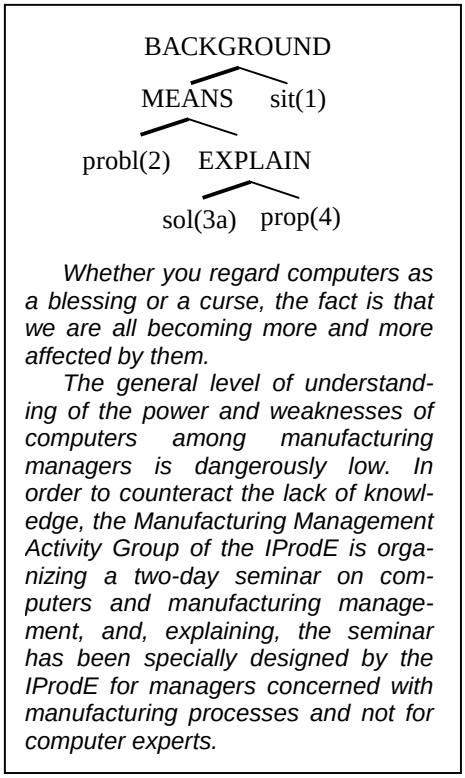


Figure 7: Plan 2

4. Assessing DMSumm

We followed several approaches (Sparck Jones and Galliers, 1996; White et al., 2000; Mani, 2001) to assess DMSumm in the light of its three posed constraints. We used a test corpus in Portuguese (Feltrim et al., 2001), hereafter named Thesis Corpus, composed of 10 postgraduate introductions (of MSc. monographs and PhD. theses) in Computer Science, for the experiments described here.

4.1. Experiment 1: Identifying CPs

Nine judges, computational linguists and native speakers of Portuguese, were given the Thesis Corpus and asked to select a sentence that best mirrored each source text, assuming that just one, the most voted, sentence would signal that source text CP.

Most judgments (60%) indicated *solution* as the CP; the others related it to *method*, *problem*, *result* and *general proposition* (only once, for the remaining source texts). This makes evident that 1) the P-S super-structure played its role in determining texts centrality; 2) since

indicative phrases were usually explicit in the pinpointed sentences, they may have been the basis for most judgments, reinforcing previous work on extracting the main topics of a text (e.g., Paice, 1981; Hoey, 1983). We used such results to define our CPs in Experiment 2.

4.2. Experiment 2: Assessing Summaries

Ten judges, also computational linguists and native speakers of Portuguese, were given the very same Thesis Corpus, along with 4 summaries for each source text – 3 automatic and the authentic one. The automatic ones corresponded, in average, to 40% of the source texts length and were completely produced by DMSumm. The judges' task was twofold: to identify the authentic summary (Task 1) and to score each summary according to two decision points: textuality and gist preservation, considered altogether (Task 2). The score range is shown in Table 2. In most cases, the judges readily identified correctly the authentic summaries in Task 1. This can be justified by the fact that such summaries usually convey information that not necessarily appears in the corresponding source texts (due to the author's background knowledge) and/or have a richer syntactic structure than the automatic ones.

Table 2: Scores for assessing summaries

Textuality	Gist	Score
Kept	Preserved	6
Damaged	Preserved	5
Kept	Partially Preserved	4
Damaged	Partially Preserved	3
Kept	Not Preserved	2
Damaged	Not Preserved	1

Chart 1 synthesizes the judges' evaluation in Task 2, showing the average scores for automatic and authentic summaries[2]. Following White et al. (2000), the *means* measure, of the judges' answers, was calculated for such average scores, yielding a baseline score 3 for satisfactory summaries. As we can see, most automatic summaries are acceptable, when we consider textuality and gist preservation. Actually, only two of them are below the baseline. Moreover, they have been judged quite similarly to the authentic ones, in what concerns most of their corresponding scores distances, which show very few quality discrepancies. Promising results are also achieved when we consider the distribution of automatic summaries judgments per score (84% above the baseline): 25% were given the score 6; 6% the score 5; 31% the score 4; 22% the score 3; 7% the score 2; 9% the score 1[3].

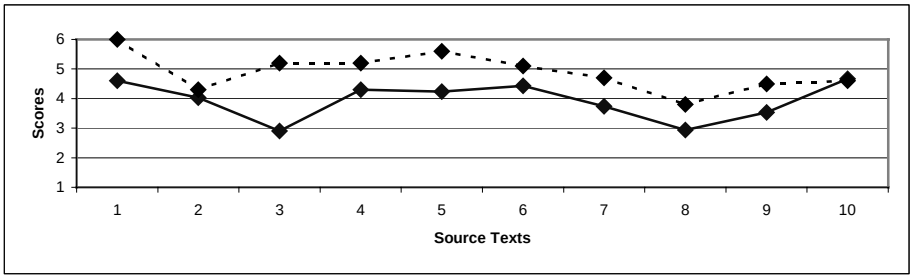
Now, if we consider only textuality as a good indicator of summary quality (Mani, 2001), the majority (67%) of automatic summaries were considered to keep it. However, the corresponding rate cannot be considered satisfactory, when compared to the judgment on the authentic summaries (90% were considered *textual*). Such a low percentage is partly due to the performance of the linguistic realizer, but it may also be the case that assuming CP as conveyed by a unique information unit is a too strong constraint to assess textuality. Actually, most

[2] Throughout this paper, solid lines signal automatic summaries and dotted lines, authentic ones.

[3] However strange it may be to include the baseline in the acceptance level, we stress that we follow White et al.'s approach. This certainly should be reviewed in further work.

often sparse information is also important in asserting a CP, but this is not considered in our experiment. Such a problem is also evidenced by the gist preservation distribution of the automatic summaries: 61% of the automatic summaries preserved only partially the gist and 31% totally preserved it. In contrast, all the authentic summaries preserved gist, according to the judgments.

Chart 1: Average scores of all the summaries per source text



Mani still suggests that a readability score can indicate how close to their authentic summaries and corresponding source texts the automatic ones are. Chart 2 shows the readability scores for the summaries and the source texts (dotted lines with square markers), following the Flesch scores, now adapted to Portuguese (Martins et al., 1996). The bigger the score, the easier it is to read the text (interval 25-50 signals difficult texts; 0-24, very difficult ones).

Experiment 2 also allowed for a comparison between summaries, by measuring their semantic informativeness, i.e., the amount of content information that is conveyed by both the summaries and related source texts (Mani, 2001). The higher the score (maximum=1), the more informative the summary. Chart 3 shows the average scores of our summaries for each source text. The biggest discrepancy, when comparing the scores of the authentic summaries with the corresponding automatic ones occurs only once (source text 3).

A comparison between the automatic and authentic summaries has also been carried out. In this case, our ideal solutions are the information units present in the authentic summaries. DMSumm performance yielded the following rates for precision, recall and f-measure, respectively: 44%, 54%, and 48%. Although authentic summaries can present more information than the source text, this problem was minimally detected in the Thesis Corpus and did not alter significantly the obtained results.

Chart 2: Readability scores

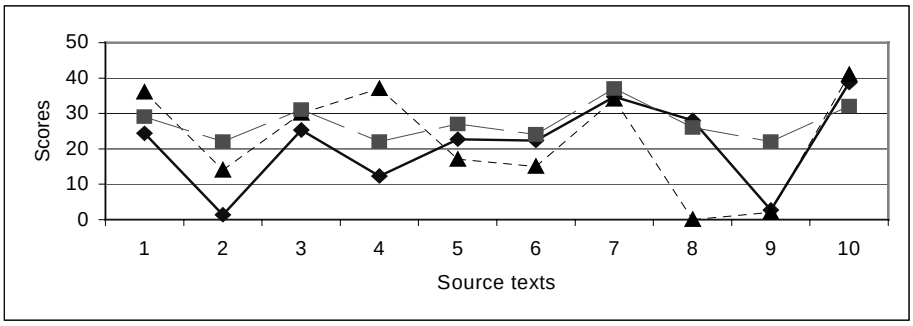
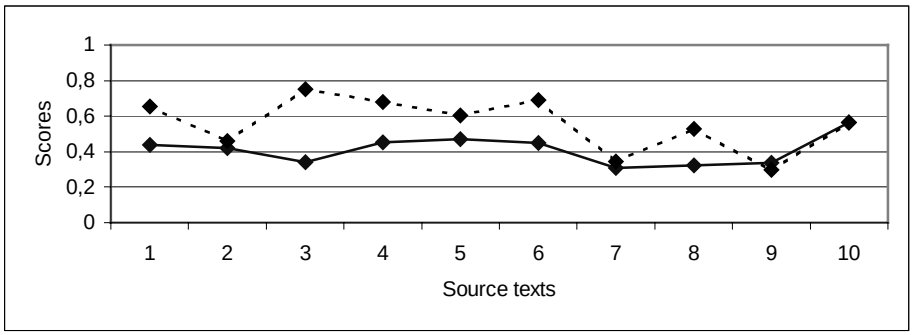


Chart 3: Semantic informativeness scores of all the summaries



5. On DMSumm potentialities

It is evident that DMSumm is actually only a text generator that reasons on summarizing constraints. For this reason, we assume that its input message is consistent with the related source texts and proceed in assessing DMSumm as if the automatically produced summaries were corresponding to source texts instead. Our assumption proved harmless by the results shown in the previous section. These make evident that the automatic summaries do not present significant coherence problems and are close to the source texts. Moreover, gist is preserved, in spite of the CP being only partially kept in most cases, since gist is also conveyed by satisfying the CG. It seems to us that CG satisfaction is quite straightforwardly guaranteed due to the technicality of our corpus. However, further exploration is needed for other text genres.

With respect to the AS task, DMSumm is deeply dependent upon the P-S account, adding an extra level of complexity to DMSumm and, actually, consisting of its main bottleneck. Given that DMSumm has so far referred to text generation, such a problem is mostly significant under text planning, which is quite complex due to its strategies based on three different types of discourse knowledge. Actually, both intentions and semantics refer back to the P-S setting. Since semantics is hypothetically addressed at the input, message, level, this should not be the focus of the current DMSumm improvement. Our question is, thus, how to get rid of the P-S dependency with no prejudice of its input. This implies questioning the need for the intentional level. Our first attempt was to suppress such a representation level and run DMSumm again. Although the automatic results showed that there was a substantial performance decrease, the results are still inconclusive and further investigation shall be pursued in the future.

References

Feltrim, V.D.; Nunes, M.G.V.; Aluísio, S.M. (2001). *Um corpus de textos científicos em português para a análise da estrutura esquemática*. Série de Relatórios Técnicos do NILC, NILC-TR-01-4. São Carlos – SP. Brazil.

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

Hobbs, J.R. (1985). *On the Coherence and Structure of Discourse.* Tech. Rep. CSLI-85-37, Center for Study of Language and Information, Stanford University.

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

Jordan, M.P. (1980). Short Texts to Explain Problem-Solution Structures – and Vice Versa. *Instructional Science*, Vol. 9, pp. 221-252.

Jordan, M. P. (1992). An Integrated Three-Pronged Análisis of a Fund-Raising Letter. In W. C. Mann and S. A. Thompson (eds.), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pp. 171-226. John Benjamins Publishing Company, Amsterdam.

Jordan, M,P. (2000). A Pragmatic/Structural Approach to Relevance. *Technostyle*, Vol. 16, No. 2, pp. 47-67.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

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

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD. Thesis. Department of Computer Science, University of Toronto, Canada.

Martins, T.B.F.; Ghiraldelo, C.M.; Nunes, M.G.V.; Oliveira Jr., O.N. (1996). *Readability Formulas Applied to Textbooks in Brazilian Portuguese*. Notas do ICMSC-USP, Série Computação. São Carlos – SP, Brazil.

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

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

Paice, C. D. (1981). The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. *Information Retrieval Research*. Butterworth & Co.

Pardo, T.A.S. and Rino, L.H.M. (2001). A summary planner based on a three-level discourse model. In the *Proc. of the 6th NLPRS – Natural Language Processing Pacific Rim Symposium*, pp. 533-538. National Center of Science, Tokyo, Japan. 27–29 November.

Rino, L.H.M. (1996a). *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. Tese de Doutorado. Universidade de São Paulo, Brasil.

Rino, L.H.M. (1996b). A sumarização automática de textos em português. In *Anais do II Encontro para o Processamento Computacional de Português Escrito e Falado*, pp. 109-119. Curitiba - PR. Outubro.

Rino, L.H.M. and Scott, D. (1994). *Automatic generation of draft summaries: heuristics for content selection*. ITRI Tech. Report ITRI-94-8. University of Brighton, England.

Sparck Jones, K. and Galliers, J. R. (1996). Evaluating Natural Language Processing Systems. *Lecture Notes in Artificial Intelligence*, Vol. 1083.

White, J.; Doyon, J.; Talbott, S. (2000).Task tolerance of MT output in Integrated Text Processes. In the *Proc. of the Embedded MT Systems Workshop*, pp. 9-16 (NAACL-ANLP 2000 Workshop). Seattle, WA.

Winter, E.O. (1977). A Clause-Relational Approach to English Texts: A Study of Some Predictive Lexical Items in Written Discourse. *Instructional Science*, Vol. 6, No. 1, pp. 1-92.