University of Brighton

Automatic generation of draft summaries: Heuristics for content selection

Lucia Helena Machado Rino
Donia Scott

ITRI-94-8

August 1994

Information Technology Research Institute Technical Report Series
ITRI, University of Brighton, Lewes Road, Brighton, East Sussex BN2 4AT

Automatic generation of draft summaries: heuristics for content selection[1]

Lucia Helena Machado Rino[2]
and
Donia Scott
Information Technology Research Institute
University of Brighton
Lewes Road, Brighton BN2 4AT
UK
Lucia.Rino,Donia.Scott@itri.bton.ac.uk

Abstract

We report on work aimed at providing discourse strategies for the automatic generation of draft scientific summaries. We have developed a discourse model based upon a theory of discourse structure and an analysis of data corresponding to available physics abstracts and naturally produced summaries written by domain-expert writers.

In this paper, we present the results of our analyses and outline some derived heuristics for compression of discourse structures that convey a selected degree of detail. When applied to the data, the heuristics provided coherent and clear summaries, significantly shorter than the originals. These heuristics guide the process of compression in the automatic summary generator that is currently under development.

Introduction

The work described in this paper is part of a project for the automatic generation of summaries of scientific papers in English. Summarisation involves, among other things, the selective choice of key information units to convey to the user.

In this paper, we will report on some of the techniques we have developed for driving automatic content selection. Discourse strategies are used to select and organise information for the generation of different versions of summaries. Only content units that are significant for a particular readership must be expressed in the generated summaries, where by significant we mean being understandable and relevant for the reader's purposes. The versions differ on parameters of informativeness and conciseness, for different communicative purposes.

[1]Paper presented at the Third International Conference on the Cognitive Science of Natural Language Processing. Dublin City University, Ireland, July 1994.

[2]On leave from the Universidade Federal de São Carlos - SP - Brazil. This work is supported by the National Council for Scientific and Technological Development (CNPq), Project No. 201610/92-2, and the Fapesp Project No. 92/2151-8.

The source of information for the generation of summaries is a set of content units (e.g. propositions), ranked according to their importance. They can be selected in different ways, depending on the communicative goals of the summary to be generated. For example, comparing informative and indicative summaries of the same text (i.e. those which present substantive information of the corresponding paper versus those which signal the content of the corresponding text but do not refer to substantive information (Hutchins, 1987; Paice, 1990)), the kind of information conveyed differs substantially, although the message may be the same. In these two types of summaries, the degree of informativeness is the main variable factor. Informative summaries convey most information available, whilst indicative summaries only signal this information.

Different discourse strategies can apply to distinct goals related to informativeness, so that a wide range of summaries can be produced. The transmitted message must, however, be always clear and coherent. To investigate the phenomenon of compression and informativeness, we start with a discourse structure of an informative summary, and analyse how optional content units can be suppressed.

A discourse model has been developed through the study of corpora in the domain of physics. In a first phase of the investigation, informational content and macro-structure organisation have been drawn from physics abstracts published in technical journals. In a second phase, a corpus has been collected via an empirical study consisting of a summarisation task carried out by domain-expert writers. Both experiments have provided similar data concerning informational content and macro-organisation. The analysis of the naturally produced summaries suggests that different summaries convey content at different levels of representability, by means of diverse discourse structures. We will present some of the derived heuristics for guiding content selection for different types of summaries. Throughout we will illustrate the operation of the heuristics with respect to the following naturally produced summary, shown here with clause delimitation:

Text 1: Original summary delimited by clauses[3]

1. The interaction of continuous CO2 laser radiation focused onto a free water surface is studied, both in normal gravity and in reduced gravity conditions.
2a. The [2b] depth of the keyhole structures produced by different laser powers are found to be in good agreement with the theory of Andrews and Atthey.
2b. observed
3a. This theory includes the recoil pressure of the evaporation, hydrostatic pressure and surface tension
3b. but does not include dynamic effects,
3c. such as the momentum reaction flow.
4. The shape of the keyhole and variation of depth with gravity are also calculated.
5a. The size distribution of bubbles produced at the tip of the keyhole has been measured
5b. and the [5c] is explained by the increasing sharpness of the tip.
5c. observed trend towards smaller bubbles at higher power
6a. Using analysis of the balanced forces on bubbles trapped under the keyhole,
6b. the speed of the momentum reaction flow down the side of the keyhole has been calculated to be about 20 cm/s.
7. This is a significant flow which has not been considered in previous theoretical

[3]Clauses represented in square brackets appear in the natural flow of the sentence, and they are generally expressed as adjectives, or adjectival clauses.

models.
8a. Very large bubbles have been observed during a transition into low-gravity conditions
8b. and are partially explained in terms of the pressure difference between the narrow keyhole and the initial bubble.
9. It is suggested that this effect may be of significance in laser-beam welding.

(217 words)

Analysis of physics abstracts occurring in technical journals

In the first stage of the investigation, a corpus of one hundred physics abstracts of highly specialised magazines has been selected and analysed[4]. The analysis has been carried out to determine the most common components of the abstracts, according to three perspectives: (1) the macro-structure; (2) the most relevant content information, and (3) the relationship between abstracts and corresponding source texts (to verify if the structures of the abstracts were actually *mirrors* of the corresponding source structures).

The results show that the macro-structure of the abstracts corresponds to an *Introduction-Methodology-Results-Discussion-Conclusion* sequence (Weissberg and Buker, 1990), in which the introduction very often embeds background information and the aim of the research, and the conclusions, a discussion about the investigation carried out or the results achieved. This structure suggests that the topic information of each important section of the corresponding source texts is selected to compose the abstracts. In other words, the abstracts mirror the texts.

The informational content in the corpus includes research topic, materials used and procedures adopted in the investigation, conditions under which the investigation is carried out, results achieved, practical application of the results, theoretical framework of the research, and others. We will refer again to such an informational content in this paper, since the second experiment addresses this issue in a more detailed form.

An important conclusion of this analysis is that physics abstracts convey a high proportion of indicative features. This is partly due to the high level of condensation of such abstracts (in general, they are approximately one hundred fifty words long), and to the high degree of technicality of this genre of discourse. However, there is no clear evidence that highly indicative summaries such as the ones analysed are suitable for the readers in that community. We performed a second experiment in order to obtain a richer corpus of physics summaries and analyse the interplay between selection, organisation and expression of information.

Analysis of naturally produced physics summaries

In this second experiment, subjects (divided in two groups: domain-experts and -novices) write two short summaries of the same technical paper, where each summary is aimed at a different assumed readership. Concerning content selection, the full set of units of information corresponds to the same data found in the previous experiment. However,

[4]It is worth noting that we have not constrained the type of summaries analysed, i.e. both theoretical and experimental work have been considered.

there are clearly some concepts that are more frequent than others . Our results reveal a core informational content, identifiable in the summaries written by both groups (see Table 1). This core can give rise to obligatory or optional content units. The former will compose every generated summary, whilst the latter will be selected for omission, or inclusion in the summaries.

Information content	**Degree of representability**
Research topic	80%
Materials used in the investigation	80%
Procedure adopted in the investigation	85%
Conditions of the experiment	100%
Results achieved	100%
Practical application of the results	80%
Indication for future work	80%
Comparison with other work	80%
Topic of the paper	70%
Explanation of the process	85%

Table 1: Degrees of representability of informational content in the corpus

In the linguistic analysis, we focus on theories of discourse structure for the representation of the discourse under investigation, by means of global discourse structures that can legitimate automation (Hoey, 1983; Hutchins, 1977; Hutchins, 1987; Kintsch and van Dijk, 1978).

Recent research in computational linguistics, and particularly in summarisation, have been pursuing similar goals for particular genres and styles of text (Endres-Niggemeyer, 1993; Liddy, 1991; McKeown, 1993; Paice, 1990; Sparck-Jones, 1993). Much of the work developed so far for summarisation addresses discourse modelling by means of analysing corpora linguistically. Some still suggest schematic representations of discourse strategies (Liddy et al., 1993; McKeown, 1985; Paris, 1993) for the generation of texts.

In our work, schematic representations of discourse strategies are also derived for content selection and organisation of information. We adopt the pattern of scientific discourse known as Problem-Solution (P-S) (Hoey, 1979; Hoey, 1983; Hutchins, 1977). According to this pattern, at a macro-structural level the summaries express the sequence

Situation-Problem-Solution/Response-Results-Evaluation

which is corresponding to the Introduction-Methodology-Results-Discussion-Conclusion sequence previously assigned to the first corpora. These macro-components can be further detailed in terms of other macro-components. For example, each result can be independently evaluated, or a problem can be stated by highlighting the negative aspects that need to be investigated. They can be combined in a semantic progression of discourse segments (Hutchins, 1977), in a very intricate logical sequence, according to the 1. Here, possible sequences are roughly represented ("Begin" and "End" tags highlight initial and

final structure organisers, respectively[5]). Lateral information corresponds to discourse segments that hold at intermediate levels of discourse organisation. The central sequence represents the highest level of discourse organisation, i.e. the macro-structure.

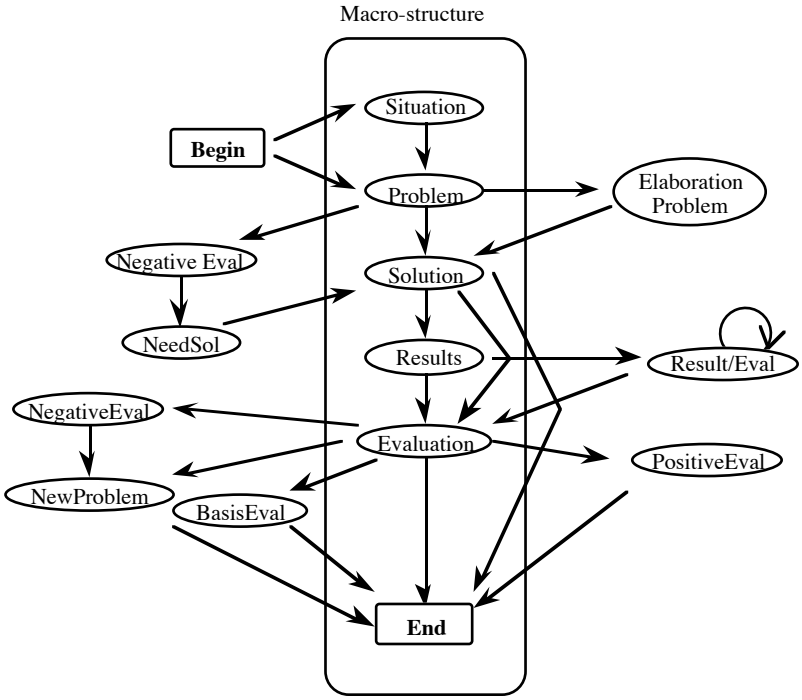

Figure 1: The Problem-Solution structure

According to this paradigm, it is possible to build infinite patterns of discourse from a finite number of resources. Thus, such a pattern is also suggestive for generation purposes.

Discourse relations in the corpus items

Discourse relations are assumed to exist between information units corresponding to clauses or groups of clauses. They characterise text structure at all levels of discourse representation (Hobbs, 1978; Mann and Thompson, 1987). For example, a Cause-Consequence (Ca-Co) relation holds between the statement of a Problem and the statement of a Solution: "Because there is a problem, a solution is searched for."

In the analysed corpus, the discourse relations suggested in the Problem-Solution pattern hold. Building on this work, we also make use of other discourse relations to express the links between discourse segments at more fine-grained levels. For example, we include in our framework relations such as Purpose, Enablement and Justification (Mann and Thompson, 1987), Means (Moore, 1989); Exemplification and Explanation (Hobbs, 1985); Background, Evaluation and Elaboration (Hobbs, 1985; Mann and Thompson, 1987). Relations that are too abstract for a good characterisation of the discourse organisation are further detailed. For example, relations that introduce particular information

[5]The minimum path comprises only the discourse components Problem and Solution.

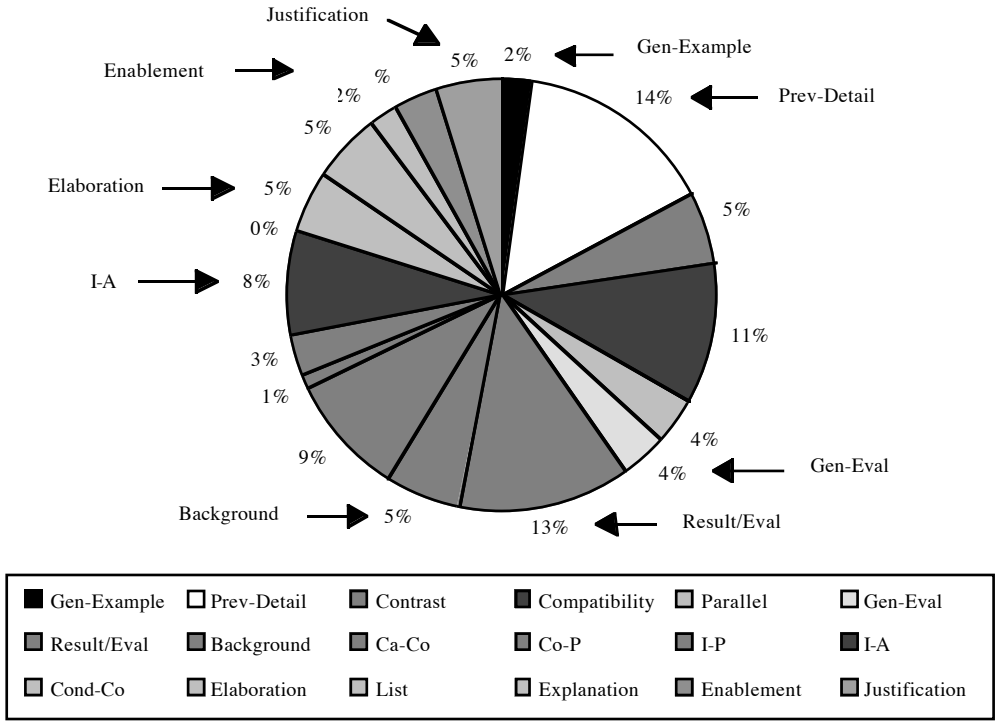

Figure 2: Discourse relations and frequency in the corpus

(i.e. General-Particular relations) can be expressed by General-Example and Preview-Detail relations. Details of the necessary modifications will not be reported here.

A full account of the extracted discourse relations, and their frequency in the corpus, is given in Figure 2. The relations highlighted in this figure guide the definition of the heuristics for compression presented in this paper.

Definition of heuristics for compression of a discourse structure

We adopt the following assumptions for compression in summary generation:

- In the discourse structure under investigation, there are non-essential units of information that can be omitted.
- Whenever omitting a component from a discourse structure, the resulting text is still understandable, at least for more knowledgeable readers (i.e. those able to bridge the resulting inference gaps).
- In omitting discourse components, all the possible referents are resolved in advance.

- The deletion of complex discourse segments (i.e. those that do not relate only propositional-like units) implies the deletion of the related sub-structures.
- The suppression of intermediary relations of a discourse structure does not necessarily imply the suppression of its macro-components.
- Omitting all the "possible" (i.e. not "necessary") discourse components of a discourse structure leads to a highly indicative summary.
- The same discourse component that can be omitted in one context may be obligatory in another, depending on the addressed readership[6].
- The degree of informativeness/indicativeness of a summary can vary. Different heuristics can be applied to the same discourse structure, once they are proved to be compatible.
- The quality of the summaries resulting from the application of some heuristics is not evaluated at this stage of the research. However, coherence must be assured, even though it may not be explicitly marked by cohesive devices.

Compression can happen in two ways: (a) by considering the suppression of macro-components of the discourse structure; (b) by considering the suppression of information related to the expression of more detail (which features specific discourse relations).

In the case of (a), eliminating a macro-component of the discourse structure implies in eliminating all its derived sub-structures (this is one of the assumptions).

In the case of (b), eliminating details involves analysing which relations that hold in the intermediate level of discourse organisation can be suppressed, with no prejudice to coherence. More detail is directly linked to General-Particular (which introduce exemplifications and details of discourse entities) and Matching relations (which introduce comparisons and parallelisms between discourse entities) in the P-S paradigm. Other relations, such as Elaboration, which can be further refined as Part-Whole, General-Specific, Abstract-Instance, and Attribute (Hovy, 1990), also allow compression (they can be used for generalisation, which is a widely used device for summarisation).

In what follows, we outline some of the heuristics we have developed for content selection and give examples of their application with reference to the naturally produced summary shown in Text 1. Each heuristic is derived from a hypothesis arising from the results of the corpus analysis.

Heuristic 1: Delete Particular, from General-Particular relations.

Applying this heuristic to clauses 3b-3c of Text 1, given by:

> `[This theory]` (*resolution of the reference to clause 3a*) `does not include dynamic effects, such as the momentum reaction flow.`

the resulting sentence is

> `This theory does not include dynamic effects.`

[6]Although we are not addressing in this paper the selection of content according to the readership, such an assumption allows different content selection to take place for different versions of summaries.

Heuristic 2: Delete every discourse segment that is optional, and linked at any level to Problem-Solution segment[7].

When this heuristic is applied to the entire summary, the result is the minimum summary

> `The interaction of continuous CO2 laser radiation focused onto a free water surface is studied, both in normal gravity and in reduced gravity conditions.`
>
> (24 words)

Heuristic 3: Delete Y, in *Elaboration(X,Y)*

Applying this heuristic to the sentences 2-3 of Text 1, given by:

> `The observed depth of the keyhole structures produced by different laser powers are found to be in good agreement with the theory of Andrews and Atthey. This theory includes the recoil pressure of the evaporation hydrostatic pressure and surface tension but does not include dynamic effects, such as the momentum reaction flow.`

the resulting sentence is

> `The observed depth of the keyhole structures produced by different laser powers are found to be in good agreement with the theory of Andrews and Atthey.`

Other important heuristics include:

Heuristic 4: Delete Results, in *I-A(Solution,Results)*.
Heuristic 5: Delete Y, in X *Evaluation(X,Y)*.
Heuristic 6: Delete Solution, in Solution-Result, when it is an intermediate discourse segment.
Heuristic 7: Delete Y, in *Justification(X,Y)*.
Heuristic 8: Delete Detail, in a Preview-Detail relation.
Heuristic 9: Delete Example, in a General-Example relation.
Heuristic 10: Delete X, in *Background(X,Y)*.
Heuristic 11: Delete Ca in a Ca-Co-P relation, when Purpose (P) is a repetition of Cause (Ca).
Heuristic 12: Delete X, in *Enables(X,Y)*.

A clearly careful analysis of the coverage of the outlined heuristics and hypotheses is required for integration into discourse strategies for generation of draft scientific summaries.

[7] A similar method has been suggested for summarisation purposes: if we consider only the most significant units of information linked to the macro-components (e.g. the topic information, or the most nuclear, in rhetorical terms), the result will be a minimum skeleton, corresponding to a highly indicative summary (Hoey, 1983).

Discussion

We follow a tradition in discourse processing, which has been strongly influenced by work of Winter (Winter, 1979) and Hoey (Hoey, 1979; Hoey, 1983). In computational linguistics, important work in this field has been carried out by describing a text structure as a tree of relations holding between pairs of spans of text (Hobbs, 1985; Mann and Thompson, 1987). Schema-based approaches have also oriented discourse organisation through combinations of rhetorical predicates (e.g., (McKeown, 1985; Paris, 1993)).

In our work, we integrate different types of discourse relations in order to organise discourse coherently at a macro-level. We consider schemata appropriate for a precise account of coherence, and thus, for the production of suitable and understandable summaries[8]. Different schemata can correspond to different discourse strategies, so that different purposes of communication can be formulated and expressed to compress/expand discourse structures. Coherence can, thus, be enforced by the correct nesting and filling of schemata (Hovy, 1990).

The data provided by our linguistic analysis show that an average of seventeen relations hold in the corpus. From these, an average of 36% correspond to General-Particular or Matching relations; others equally important for compression are Enablement, Background, Evaluation, Justification, and Elaboration. These account for an average of 35% of the relations in the corpus. Altogether, these relations signal possibilities for omission of information. The "trial" set of heuristics based on them seems quite promising. The resulting summaries appear satisfactory: they are more indicative and significantly more condensed than the originals, and still convey coherent messages. Meaning is preserved, to the extent that it is possible to consider preservation of meaning in indicative summaries.

In this work, we deal only with structural features. For a complete generation framework, other heuristics that lead to concise, but still informative summaries, must be specified for a right balance between conciseness, informativeness and clarity. We are about to incorporate and implement the heuristics presented in this paper into discourse strategies for the generation of draft scientific summaries. Lexicalisation and grammaticalisation of the discourse structures will be investigated in a later stage.

References

Endres-Niggemeyer, B. A Naturalistic Model of Abstracting. In Dagstuhl Seminar Report (9350), *Workshop on Summarising Text for Intelligent Communication*. Dagstuhl, Germany, 1993.

Hobbs, J.R. *Why is Discourse Coherent?* Tech. Note 176. SRI International, California, 1978.

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.

[8]It is worth noting that, although schema-based approaches seem insufficient as a discourse model for certain types of discourse (see, for example, (Moore and Paris, 1993)), they are not problematic for the generation of summaries.

Hoey, M. *Signalling in Discourse*. Discourse Analysis Monographs, English Language Research. University of Birmingham, 1979.

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

Hovy, E. Unresolved Issues in Paragraph Planning. In R.Dale; C.Mellish and M. Zock (eds.), *Current Research in Natural Language Generation*, pp. 17-45. London, Academic Press, 1990.

Hutchins, J. *On the Structure of Scientific Discourse*. UEA Papers in Linguistics 5, pp. 18-39. University of East Anglia, 1977.

Hutchins, J. Summarization: Some Problems and Methods. In K.P. Jones (ed.), *Meaning: the frontier of informatics*, pp.151-173. Aslib, London, 1987.

Kintsch, W. and van Dijk, T. Toward a Model of Text Comprehension and Production. Psychological Review Vol. 85, No. 5, pp. 363-394, September 1978.

Liddy, E.D. The Discourse-Level Structure of Empirical Abstracts: An Exploratory Study. *Information Processing & Management* 27(1), pp. 55-81, 1991.

Liddy, E., McVearry, K.A., Paik, W. Yu, E. and McKenna, M. Development, Implementation and Testing of a Discourse Model for Newspaper Texts. In *ARPA Human Language Technology Workshop Proceedings*. N.J., 1993.

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

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

McKeown, K.R. Generating the Complex Sentences of Summaries Using Syntactic and Lexical Constraints: Two Applications. In Dagstuhl Seminar Report (9350), Workshop on *Summarising Text for Intelligent Communication*. Dagstuhl, Germany, 1993.

Moore, J.D. *A Reactive Approach to Explanation in Expert and Advice-Giving Systems*. PhD. Dissertation, University of California, 1989.

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.

Paice, C.D. Constructing Literature Abstracts by Computer: Techniques and Prospects. *Information Processing & Management* 26(1), pp. 171-186, 1990.

Paris, C. L. *User Modelling in Text Generation.* Pinter Publishers, 1993.

Sparck Jones, K. What might be in a summary? In Knorz, Krause and Womser-Hacker (eds.), *Information Retrieval 93*, pp. 9-26. Universitatsverlag Konstanz, June 1993.

Weissberg, R. and Buker, S. *Writing Up Research: Experimental Research Report Writing for Students of English*. Prentice Hall, Inc., 1990.

Winter, E.O. Replacement as a Fundamental Function of the Sentence in Context. *Forum Linguistics*, Vol. 4, No. 2, pp. 95-133, 1979.