How to Appreciate the Quality of Automatic Text Summarization? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN

Jean-Luc Minel *, Sylvaine Nugier ** and Gérald Piat ***

* CAMS-CNRS
96 Boulevard Raspail
75006 Paris
minel@cams.msh-paris.fr

** EDF DER/GRETS
1 Avenue du Général de Gaulle
92000 Clamart
[Sylvaine.Nugier, Gerald.Piat@der.edfgdf.fr

Abstract

For the SERAPHIN project, we set up two assessment protocols in order to be able to more accurately assess the quality of abstracts - the FAN protocol and the MLUCE protocol, for which we provide the results The FAN protocol assesses the legibility of an abstract, independently from the source text The MLUCE protocol is designed to allow users of automatic abstracts to assess their quality These protocols were applied to a corpus of 27 texts which varied in length from between three and twelve pages These texts were randomly chosen from EDF archives They include both scientific and general press articles, extracts from books, and internal EDF notes The results of the FAN protocol demonstrate the difficulty of using surface linguistic indicators to assess the quality of an abstract, the results of the MLUCE protocol illustrate the importance of user expectations

1 Introduction

The SERAPHIN system produces abstracts using an alternative approach, the contextual exploration method (Desclés et al 97), based on the pinpointing of linguistic indications in order to identify i) certain structuring information, ii) causal arguments and arguments by cause (Jackiewicz 96), iii) different defining wordings (Cartier 97)
The abstracts are made up of sentences extracted from the source text, and to which semantic labels have been attached (Berri et al 96), representing the salient points of the source text from the author's point of view The size of the abstract is limited to 20% of the source text
The assessment of abstracts has often been approached from the computer documentation angle (Salton 89), particularly by using criteria such as system's recall and system's precision The main problem with these criteria is the postulation of the existence of a user request, expressed in the form of a combination of describers It can be seen that this hypothesis does not generally correspond to the reality of using abstracts

Indeed, the reader of the abstract has already selected the source document as being one which belongs to his field of interest, and he is looking to obtain the most accurate understanding possible of the content of the document This is why, within the SERAPHIN project, we set up two assessment protocols, FAN and MLUCE, which are designed to better assess the quality of abstracts The purpose of the present article is not the evaluation of the SERAPHIN system itself[1], but "how to evaluate" the quality of automatic text summarization system

At the outset, we wished to assess 50 texts, but the cost, in terms of reading time, forced us to reduce this objective These two protocols were applied to a corpus of 27 texts which varied in length from between three and twelve pages These texts were randomly chosen from EDF archives They include both scientific and general press articles, extracts from books, and internal EDF notes

2 The FAN Protocol

This protocol aims to assess the quality of an abstract independently from the source text and the information it contains Assessment was therefore carried out by two jurors, who were not specialists in the fields concerned, who read the 27 abstracts without having seen the source texts It takes approximately 10 minutes to assess an abstract The assessment grid contains 4 criteria, which are described below

Criterion 1 Number of Anaphora Deprived of Referents

Given that an abstract is created from sentences from the source text, it is possible that a sentence contains an anaphora whose referent does not belong to the extracted sentence We must not forget that SERAPHIN does not detect referents, it detects what it considers to be indications of potential anaphora within sentence P1, out of a closed list (*it, this, that*, etc), and then applies a simple heurism which involves selecting the preceding sentence P0 which contains the potential anaphora On the one hand, this heurism may prove to be insufficient (no explicit referent, the referent is located in a sentence further up in the text), and on the other hand this heurism is not applied to sentence P0 (in order to avoid selecting sentences based on criteria that are

[1] this will be the object of another report

not "semantic") We believe this criterion to be a determining factor with regard to the legibility of the abstract, nevertheless the jurors raised several problems We will illustrate these via a number of examples

Thus, in the following sentence,

According to this researcher, it is a movement which is characterised by a desire to go backwards, a desire for a state in which the distinction between subject and object no longer exists (Text N°2)

the anaphoric term is deprived of its referent but the legibility and coherency of the abstract are not really altered, because, in this textual context, the name of the researcher is not considered to be an important piece of information Nevertheless, in order to restrict the effects of interpretation, this case was considered to be an anaphora deprived of a referent

In the following sentence,

One could say that these first contacts between EDF and the resident created a precedent which was not very favourable with regard to establishing peaceful relationships between EDF and the local population living near the line (Text N°9)

the potentially anaphoric term *these* may be interpreted as referring either to contacts described earlier in the text, or to a chronological account which takes on meaning as one reads the entire text, and, in part, at the end of the sentence in question In the text under consideration, it is the second interpretation which is correct, but because the juror did not have access to the source text, we treated this case as an anaphora deprived of a referent

Criterion 2 Rupture of textual segments organised by Linear Integration Markers

Various studies on textual linguistics (Charolles 89, Adam, 90) have underlined the utility of locating linguistic markers in order to identify the discursive organisations which go beyond the sentence itself SERAPHIN identifies linear integration markers (MIL) from within a closed list (*on the one hand, on the other hand, firstly, secondly, etc*) in order to rebuild textual segments in the abstracts produced Thus, if sentence P is selected, sentences P_1, P_2, P_n which are linked to P via MIL's will also be selected This selection may fail, either because the MIL is not recognised (absent from the list, ellipsis, spelling mistake, etc), or because the maximum size of the abstract has been reached We should stress both the fact that the jurors can only detect ruptures in the textual segments, and not their completeness, and that argumentation connectors (such as *indeed, furthermore*, etc) are not considered to be MIL's This decision was taken after a preliminary validation by ten or so readers, and was confirmed by the jurors SERAPHIN uses a special symbol *[ ]* to show that two sentences are not adjacent in the source text, such as in the following example

It is easy to imagine the huge number of texts and written documents that is produced by a company like EDF

[ ]

Of course, all "language production" may appear to come from an abstract source, a unique source (Text N°6)

This symbol avoids the problem of the reader mistakenly reconstructing argumentation chains

Criterion 3 Presence of "tautological" sentences

A sentence is considered to be tautological if the information it provides is completely independent of the source text, as in the following example

Predicting the future is a difficult and uncertain exercise (Text N°4)

We were trying to detect abstracts which, although optimal from the point of view of the two criteria above, had merely been created from very general sentences, as is often the case with certain abstracts by authors In fact, this criterion is far too dependent on the knowledge that the reader has of the subject of the text, and the results of the protocol show that it is not pertinent

Criterion 4 Legibility of the abstract

This criterion, whose values are *Very Bad, Mediocre, Good, Very Good*, is an overall appreciation of the abstract Although they are highly subjective, the "scores" given by the jurors varied very little, with just the two exceptions set out in table 1 below

	Juror J1	Juror J2
Text N° 20	Mediocre	Very Good
Text N°22	Good	Very Bad

Table 1 · Divergence of assessment between jurors

Text N°20 *(TF1, Le Grand Bluff)* is a chronological description of the privatisation of the television channel TF1 The succession of events, the number of players involved, metaphors such as *"It is at the foot of the wall that one judges the bricklayer"*, make the reading of the abstract (and of the source text) difficult unless the reader has a good understanding of the subject concerned (exceptionally, juror 2 happened to know this subject well)

Text N°22 *(Credibility of command control systems concepts and tools)* is a highly technical text outside the experience of the two jurors

For the presentation of the results (Tables 2 and 3) we systematically chose the lowest "score"

Interpreting the results of the FAN Protocol

We cross-referenced the legibility criterion with criteria 1 and 2 Contrary to our original hypothesis, there is no correlation between the two Indeed, the comments made by the two jurors show that overall reading of the abstract allows them to overcome any localised lack of understanding caused by the absence of anaphoric referents This conclusion must nevertheless be nuanced by the fact that there is a limited number of mistakes in the abstracts that were analysed Assessing abstracts simply on the basis of surface linguistic indicators, and without calling upon the knowledge that the jurors may have of the subject concerned, remains a difficult problem

No mistakes	One mistake	Two mistakes	Three mistakes

Criterion 1 (Anaphora)	13 Texts 48 %	5 Texts 19 %	6 Texts 22 %	3 Texts 11 %
Criterion 2 (MIL)	10 Texts 37 %	13 Texts 48 %	4 Texts 15 %	0 Texts 0 %
Criterion 3 ("Tautology")	25 Texts 92 %	2 Texts 8 %	0 Texts 0 %	0 Texts 0 %

Table 2 Synthesis of the results for the first three criteria

	Very Good	Good	Mediocre	Very Bad
Criterion 4 (Legibility)	7 Texts 26 %	13 Texts 48 %	5 Texts 19 %	2 Texts 7 %

Table 3 Synthesis of the results for the legibility criterion

3 The MLUCE Protocol

3.1 Objectives

The aim of this protocol is to enable potential users to assess the quality of automatic abstracts In this case, an abstract, and therefore its quality, will not be defined in any absolute way, but rather in terms of the ways in which it can be used For example, if a person is looking for the "proven" idea within a text, he will need to understand the different stages of the argument, on the other hand, if he only wishes to observe any simultaneous occurrences of two themes within the same text, he no longer needs to have the arguments The assessment of quality therefore depends on what one wishes to use the abstract for (Rath et al, 1961) It is therefore important to pre-define one or more uses of the abstract, and for each definition to accurately measure the "distance" between the source text and its abstract We selected two applications for automatic abstracts which were of particular interest to EDF

- *Application 1* the abstract is a tool which allows one to decide whether or not to read the source text
- *Application 2* the abstract is a support for writing a synthesis of a written document

The MLUCE protocol therefore aims at "measuring" how a given abstract meets these two objectives

In section 3 4, we set out the results on SERAPHIN, but this protocol was applied, without much importance, to the RAFI system (Lehmam 95)

3.2 Experiment procedure

The procedure selected for assessing a "summarising" system has to be precise, complete and unambiguous, in order to

- restrict the "reader effect" as much as possible - in other words, to limit the variation in assessment between readers, due to their different fields of expertise, different cultures or varying archetypes of abstracts,
- limit the influence of the order in which the texts are read,
- be adapted to all types of text, and to take their differences into account

A pilot was required in order to adjust the procedure to the above requirements For the SERAPHIN assessment, we set up a jury of four readers

- a qualified French language teacher, specialised in teaching abstract and synthesis techniques,
- a documentary researcher, working in a documentary unit,
- two users, with different training backgrounds

For the first stage of the assessment, we gave each member of the jury seven texts, their abstracts, and a list of instructions explaining the approach to be used (the "jurors" each had different texts Each reader-assessor then had to

- read the documents in a pre-defined order (firstly, all the abstracts, then all the source texts),
- fill in the reader's sheet (attached to each document) as he went along,
- give his overall opinion on the "comparison sheets" provided for this purpose

The second stage of the assessment involved analysing the sheets returned by the readers

The whole experiment (definition of procedure, and the actual assessment) lasted a total of eight months

3.3 Criteria retained

On the basis of the experiments carried out by Borko et al (1975), Edmunson (1969), Mathis et al (1973), Payne (1964) on the assessment of the quality of "abstracts", and in terms of the applications defined (section 3 1), we set four criteria, and for each criterion we established the means of assessing it

Application 1

For the first application, the criteria defined in MLUCE are designed to assess the utility of the abstract as a suitable decision-making tool for the reader These criteria must allow one to judge whether the abstract contains the information required to be able to decide whether or not to read the source text

In order to do so, we will say that the abstract must allow us to

- identify the field or nature of the source text Each reader fills in two grids (one for the source text and one for the abstract) which show the fields or natures of the texts scientific or technical, political, sociological, polemical, general, prospective, retrospective, situational or state-of-the-art

• check the presence of the essential ideas Each reader underlines the ideas in text T_i which he feels to be essential, and checks that they are present in abstract R_i

• avoid parasitic ideas Each reader highlights sentences in R_i which should not be in R_i, and the sentences in abstract R_i which are cut off from the context (essential ideas that have been cut short)

Application 2

For the second application, the criteria defined in MLUCE are designed to assess the utility of the abstract as a support for writing a synthesis of a written document

In order to do so, we will say that the abstract must allow us to

• identify the field or nature of the source text (criterion identical to application 1)

• check the presence of the essential ideas (criterion identical to application 1)

• highlight the logical linking of ideas Each reader fills in two grids (one for the source text and one for the abstract) in which the following argumentation links appear cause implying consequence, consequence implies cause, proposition of a solution, from particular to general, from general to particular, motivated juxtaposition of facts, listing of facts, confrontation He then states whether the idea is "proven" in each of the documents he has read Finally, he assesses whether the abstract is clear, fairly clear, not very clear or incomprehensible

3.4. Results on SERAPHIN and interpretation

Identification of the subject (table 4)

The texts submitted to the reader-assessors may cover several fields and be of several different natures, which is why the total number of texts shown in table 4 is greater than the number of texts studied (number studied = 27)

The categories listed in table 4 were not explicitly defined to the jurors, it is therefore possible that there is a certain amount of "subjectivity" in the categorisation of the texts Nevertheless, we supposed that each reader could implicitly and continuously in time divide the texts into the categories proposed

	number of source texts for which the abstract	
	respects the subject or	does not respect the subject or
scientific or technical	9	10
political	6	7
sociological	8	2
polemical	1	0
general	1	0
prospective	3	8
retrospective	2	4
situational or state of the art	15	5

Table 4 Identification of the subject

	number of abstracts
close to the text	2
fairly close to the text	11
relatively different from the text	10
well away from the text	4

Table 5 Presence of the essential ideas

Presence of essential ideas (table 5)

The result of highlighting the "essential ideas" in the source text, and of the reader marking the "parasitic ideas" that appear in the abstract, are grouped together in order to define a "proximity" indicator This indicator is defined in the following way

- we will define an abstract as being *close to the text* if more than 75% of the sentences which make it up are among the essential ideas (highlighted) and less than 10% are parasitic ideas,
- we will define an abstract as being *fairly close to the text* if between 50% and 75% of the sentences which make it up are among the essential ideas (highlighted) and less than 10% are parasitic ideas,
- we will define an abstract as being *relatively different from the text* if between 25% and 50% of the sentences which

make it up are among the essential ideas (highlighted) and less than 10% are parasitic ideas,
- we will define an abstract as being *well away from the text* in all other cases

Highlighting the logical sequence of arguments (table 6)
We have supposed that a text was written, by his author, in a precise aim (the "proven" idea) We have identified 8 types of argumentation links which allowed the authors to construct their demonstration (rows in table 6) Like when identifying a field, the texts submitted to the jurors may link several types of argument, which is why the total number of texts shown in table 6 is greater than the number of texts studied (readers were asked, where necessary, to give details of the order in which the different types appeared over the whole of the text)

	number of source texts for which the abstract	
	respects the chain of argument	does not respect the chain
cause implying consequence	4	7
consequence implies cause	1	3
proposition of a solution	6	5
from particular to general	1	2
from general to particular	0	3
motivated juxtaposition of facts	10	7
listing of facts	2	0
confrontation	1	2

Table 6 highlighting the logical sequence of arguments

Table 6 should be compared with table 4 Indeed, we have noticed, with regard to the texts studied, the absence in the abstract of the source of the argument that is in the original document Thus, a text which uses a given theory, leads to an abstract in which no theoretical base for the argument is given This might explain the bad performance of the "scientific" or "prospective" texts and the sequences of "cause implies consequence" or "consequence implies cause" types

Quality of the abstract (tables 7 and 8)
After having determined the field, the jurors noted the logical argumentation sequencing, stated the "proven" idea of the abstract, and filled in a grid in order to give their overall impression of the quality of the abstract

	number of abstracts
clear	6
fairly clear	6
not very clear	10
incomprehensibl e	5

Table 7 quality of the abstract

	number of abstracts judged to be			
	clear	fairly clear	not very clear	incompr e- hensible
scientific or technical	0	5	10	4
political	2	5	4	2
sociological	2	4	3	1
polemical	0	1	0	0
general	0	1	0	0
prospective	0	5	6	0
retrospective	1	1	2	2
situational or state-of-the-art	0	8	8	4

Table 8 quality of the abstract in terms of fields or natures of the source text

4 Conclusion

A text-by-text comparison of the results of the legibility criterion (*Very Bad, Mediocre, Good, Very Good*) in the FAN protocol, and the results of the quality of the abstract (*Incomprehensible, Not Very Clear, Fairly Clear, Clear*) in the MLUCE protocol, shows very little convergence Apart from two exceptions, MLUCE is always more demanding than FAN Here we highlight the differences in assessment between a user who reads an abstract in order to find answers to specific questions, and a reader who is not trying to assess the information content of the same abstract

The quality of an abstract depends on what the user expects from it, and only an in-situ assessment will allow one to really assess the performance of a "summarising" system Following this experiment with these two protocols, the installation of any such procedure would appear to be extremely expensive - not to mention the fact that it would require "user expectations" to be defined, and the related assessment criteria to be formalised

However, the FAN and MLUCE protocols, when applied to a test corpus which remains to be defined, may nevertheless serve as a basis on which to compare systems which summarise via sentence extraction

Acknowledgements

We would like to thank our six jurors, not only for having accepted to take part in this experiment, but also for their comments on certain aspects of the abstracts

Referencies

Adam J M (1990) , Éléments de linguistique textuelle, Mardaga, Liège

Berri, J Cartier E , Desclés, J-P , Jackiewicz, A Minel, J-L (1996) Filtrage Automatique de textes , In *Natural Language Processing and Industrial Applications*, (pp 28-35), Université de Moncton, N-B, Canada

Borko, H & Bernier, C L (1975) Abstracting concept and methods San Diego, New York, Berkeley Academic Press, 250 p

Cartier, E (1997) La Définition ses formes d'expression, son contenu et sa valeur dans les texte , thèse en cours Université de Paris Sorbonne Paris

Charolles, M (1989) Marquages linguistiques et résumé de textes , in CHAROLLES M et PETITJEAN A [éds] , *Le résumé de texte , aspects linguistiques, sémiotiques, psycholinguistiques et automatiques*, Colloque international de linguistique organisé par les Universités de Metz et Nancy II [12-13-14 sept 1989], Klincksieck

Desclés, J-P Cartier, E , Jackiewicz A Minel, J-L (1996) Textual Processing and Contextual Exploration Method In *CONTEXT 97* (pp 189-197), Universidade Federal do Rio de Janeiro, Brésil

Edmunson, H P (1969) New methods in automatic abstracting *Journal of association for computing machinery*, 16(2), (pp 264--285)

Jackiewicz, A (1996) La notion de cause pour le filtrage de phrases importantes d'un texte in *Natural Language Processing and Industrial Applications*, (pp 136-141), Université de Moncton, N-B, Canada

Le Roux, D , Minel, J L & Berri, J (1994) SERAPHIN project *First European Conference of Cognitive Science in Industry*, Luxembourg, 28-30 septembre 1994

Lehmam, A (1995) Le résumé des textes techniques et scientifiques, aspects linguistiques et computationnels, Thèse de doctorat, Université de Nancy 2

Mathis, B A , Rush, J A & Young, C A (1973) Improvement of automatic abstract by use of structural analysis *Journal of the American society for information science*, 24(2), (pp 101--109)

Mike, S Itoh, E (1994) A full-text retrieval system with a dynamic abstract generation function ,in SIGIR 94, Dublin, pp 152-161

Nugier S & Piat G (1996) Evaluation du prototype de résumé SERAPHIN protocole et premiers résultat, *Note Interne EDF*, HN-52/96/008

Payne, D (1964) Automatic abstracting evaluation support American Institute for research Pittsburgh, Penn AD 431 910

Rath, G J , Resnick, A & Savage, T R (1961) The formation of abstract by selection of sentences, Sentence selection by man and machine, The reliability of people in selecting sentences *American documentation*, 12(2), (pp 139--143)

Sabah, G (1988) L'intelligence artificielle et le langage, représentation des connaissances, Hermès C, Paris

Salton, G (1989) , Automatic text processing the transformation, analysis and retrieval of information by computer, Addison Weshley Publ Comp