A LINGUISTIC METHOD FOR TEXT FILTERING

Jawad BERRI *, Emmanuel CARTIER *, Jean-Pierre DESCLÉS *, Agata JACKIEWICZ *,, Jean-Luc MINEL ***

* *CAMS*
Centre d'Analyse et de Mathématiques Sociales
CNRS / EHESS / Université Paris 4
96, Bd Raspail - 75006 PARIS FRANCE
Tel : (1) 44 39 89 50
(berri,cartier,descles,jackiew,minel)@msh-paris.fr

***EDIAT/CR2A-DI*
9, Avenue Dubonnet - 92411 COURBEVOIE FRANCE
Tel : (1) 47 65 52 87

Abstract : The aim of the system is to provide the end-user with extracts with higher information content than the indexing string and a sufficient indication as to whether the whole text is worth reading or not. The extracts are obtained by drawing labelled sentences from the text through a linguistic method. The notion of selected sentences is discussed and the contextual exploration method is exposed. The system is already fully operational and processes heterogeneous texts; a global evaluation of the results is underway.

1 - Introduction

Recent progress in text data storage and handling seems to fall short of the expectation of end-users of conventional text-retrieval systems who are often overwhelmed by the information supplied (Salton 89). Among the various techniques proposed to solve this problem, the abstracting methods allow filtering the text data flow by providing an intermediate product between indexing strings, poorly informative and ambiguous, and the entire text, far too voluminous to be approached quickly. The increasing stream of technical and scientific literature has spurred the efforts for research in automatic abstracting since the sixties (De Jong 82; Lenhert 81; Hahn et ali 85; Sabah 88; Miike 94) but corresponding operational systems are very scarce (Le Roux 91). Several factors account for this gap between theoretical works and processing tools. The most important seems to be that all these investigations rely on the classical model of computational linguistics (going from morphological processing to pragmatic analysis) (Fuchs 93, Saint Dizier 95), bringing together important resources (dictionaries, ontology) and computational linguistic tools (parsers, grammars, conceptual representations, etc.). But the processing power of software comprising such building blocks is insufficient (Marandin 93) or at best restricted to particular types of texts at least (Pugeault 95). To bypass this obstacle, some researchers have proposed to use strictly statistical techniques (Kälgren 88, Renouf 95) or to such techniques combined with linguistic analysis (Paice 90), to extract essential sentences from original texts. These works use morphological identification, without solving ambiguity and do not make use of textual information; as a result they are confronted with dangling anaphora, coherence problems, indefinite words, etc.

In this paper we propose an alternative approach: filtering information by contextual exploration, based on linguistic clues to identify i) sentences highlighted by the author; ii) causal argumentation and argumentation by causation(Jackiewicz 97); iii) definition statements (Cartier 97). This approach

does not need any domain knowledge but requires a thorough linguistic analysis of these phenom to process heterogeneous texts.

2-The contextual exploration method

Taking into account the context to solve lexical ambiguity have been used in many applicatio (Choueka 85, Slator 91, Desclés 91). Our approach is different. On one hand, cognitive observatio of professional summarizers (Endres-Niggemeyer 96) have showed that they use textual, structur thematic and lexical markers in their search strategies. On the other, various text-centered linguist works (Roulet 85, 87, Charolles 88, 89, Adam 90) have pointed out the interest of identifying and locating linguistic markers and their combinations, to lend meaning to textual units. We hav systematised these observations by taking into account all information about textual tokens (such a contextual word meaning, word location in the sentence, sentence or paragraph location in the tex structuration level of text, graphic signs used in titles) to assign semantic labels to sentences (but no necessarily to each sentence).

For each type of text processing, contextual exploration suggests the same methodology i building a linguistic knowledge database:

- identify semantic labels which may be attached to sentences or textual units;
- identify and list relevant linguistic clues which are attached to these labels;
- write procedural steps, using formal contextual exploration rules. This is dealt with in mor detail in (Berri et ali 95, Berri 96).

3-Semantic labels

On the basis of the initial results of SERAPHIN (Le Roux 94, Berri et ali 95), we have define the following classes which regroup several semantic labels :

- identification of information regarding structure;
- identification of definitions;
- identification of causal argumentation and argumentation by causation;
- identification of theme-titles.

3.1 -Labelling structuring information

By "structuring information" we mean all sentences which indicate the theme dealt with in the following or preceding sequence. Such information provides an idea of the text, which we call a neutral summary of the original text. So far, we have distinguished three kinds of structuring information : thematic declarations (general document plan : *Dans ce document, nous étudierons les points suivants* ... or local thematic declarations :*L'oxyde nitreux est un autre gaz à effet de serre dont il convient de parler* ...), thematic recapitulatories (*Nous pouvons récapituler en disant que ..., En guise de résumé*) , and global conclusions (*Il faudrait donc .... Notre conclusion est que* ...). All these sentences are easily recognizable via combinations of linguistic markers (deictics, evaluative expressions, modals, argumentative and underlined expressions) and typographic constraints (for

example, "global conclusions" must be positioned in the final part of the text). The efficiency of this first extraction has a cognitive basis : if the author wants his reader to catch easily what is important (in his opinion) in the text, he must set typographic and linguistic hints to emphasize the corresponding parts and clearly indicate the thematic structure of his work. That is what we are looking for.

3.2 -Labelling definitions

Definitions are the second kind of extracted information. Whereas the preceding ones are "text-structuring" and constitute the core of a neutral abstract, these resort to specific needs, for example terminological extraction and/or semantic tree construction. Using classical distinctions between nominal- (N) and object-definitions (T), and usage (U) and prescriptive (P) ones, we obtain a matrix in which should hold all types of definition : (1) *La mondialisation est un phénomène qui tend à intégrer et à rendre interdépendantes les économies locales par l'élargissement des marchés au-delà des nations, principalement par le poids de plus en plus fort joué par les lobbyings financiers.* (T-U); (2) *On entend par mondialisation un phénomène qui tend à intégrer et à rendre interdépendantes les économies locales.* (N-U); (3) *j'appellerai mondialisation le processus qui aboutit à nier aux économies locales la possibilité de décider de leur propre destin par la main-mise du pouvoir financier.* (T-P); (4) *Par le terme de mondialisation, nous entendrons le processus qui tend à intégrer et à rendre interdépendantes les économies locales.* (U-P). Each type of definition is identified by specific linguistic clues. At the same time, we identify the definite term and the defining term. Note that the different types of definitions prototypically tend to define differently the definite term.

Moreover, some rules can extract the hyperonyms, synonyms and antonyms of a term (extracted from titles or looked for by the end-user) throughout the text. These rules considerably improve the definition module : it is obvious that the author can't use the same term all the time in his text, and so has to make use of coreference devices. We have so far identified some of the basic procedures of coreference, such as apposition (*L'effet de serre, processus d'emprisonnement de la chaleur dans l'atmosphère.*), phrasal nouns connected by specific conjunctions (*ou, c'est-à-dire...*), and some anaphoric restatements (*Processus*, in *ce processus...* for example, can directly refer to *l'effet de serre*, its hyponym in the preceding sequence).

3.3 -Labelling causality and causal argumentation

Our approach (Jackiewicz 1996), is based on the identification of causal data expressed within two specific contexts: causal argumentation and argumentation by causation (Perelman 1992, Plantin 1990). In the first context, the cause participates in expressing and building new knowledge (the causal link is still to be confirmed or denied). In the second one it plays the role of the argument drawing from its realistic foundation to justify choices and evaluations, or to legitimate future goals and projects.

•In causal argumentation, this context is made up of clues (conditional tense, modal verb etc.) expressing hypothetical, possible, largely proved, absolutely certain nature of causal informati present in the sentence.

Selon l'UNICEE, l'écotaxe ***aurait un effet pervers sur*** *l'économie: elle* ***entraînerait*** *un baisse de la compétitivité et de la capacité à créer des emplois.*

•In argumentation by causation, the context stretches beyond the sentence. Some clues expres assessment of consequences and others of planned actions. Causal information does not hold a centr position, it shows the possibility of these actions and how they can be implemented.

Les risques de changement climatique ***consécutif à*** *un accroissement de l'effet de serre ont conduit la France à fixer un objectif volontariste de prévention des émissions de gaz à effet de serre et à proposer un accord internationa sur les moyens de prévention.*

3.4-Labelling theme-titles.

We have not built any thesaurus or pre-established index. Nevertheless, domain-dependen information is, to a certain extent, necessary : we have resorted to an indirect procedure to capture thi kind of information, consisting essentially in phrasal nouns extracted from the title and subtitles, an proper nouns. These entities are extracted by a specific module and are considered as relevan terminological units. We have based our approach on LEXTER (Bourigault 94). As titles ar expressed in a rather rudiment syntax (particularly without any verb), some modifications have bee made to spare the categorization process needed by LEXTER. In order to extract phrasal nouns fron titles, our system applies an algorithm based on the sequentiality of border-words like prepositions pronouns, articles, and builds up a set of potential outstanding groups of words.

4 - Filtering sentences

Relying on this semantic labelling, this system is able to supply extracts answering specific end user. Sentence filtering is based on a filtering profile, selection strategies and a filtering threshold The filtering profile sets the importance of each semantic label; for example, *recapitulatory* may b more interesting, for a specific user, than *hypothesis*. Selection strategies determines how the text i investigated to select sentences; for instance one strategy begins with the introduction, then investigates the conclusion and go on with the rest of the text. The filtering threshold sets th maximun number of sentences allowed (for instance 20% of the original text).This is dealt with i more detail in (Berri et ali 96).

5-System architecture

The source-text is marked in accordance with SGML standard. Some selection rules that take into account the position of the sentence in the text and the dependence relations between various entities of the text (sections, sentences, etc.) can be used. The general architecture takes the form of a four modules written in a object language.

1) The module to identify contextual exploration clues; this module seeks contextual linguistic markers. The process has to cope with two problems : the recognition of morphological variants and the discontinuity of composite markers.

2) The sentence labelling module is implemented in the form of a task-system. Each task, applying a set of rules, gives a relevant semantic label to the sentences. A sentence may get several labels. Conflicts are solved by the following module.

3) The filtering module makes use of a filtering profile, selection strategies and a filtering threshold (see§4) to select relevant sentences.

4) The last module aims at solving consistency and legibility problems. The extract is constructed around the structure of the source-text. For every section, the system systematically takes up the titles and sub-titles as well as the selected sentences.

6 - Conclusion

At present, our system runs with 3000 markers and 150 rules. The system has been tested on about forty texts and results are promising. A global evaluation has been undertaken on a hundred texts provided by an industrial company. In order to assess the extracts thus obtained we have designed three protocols involving quantitative and qualitative criteria. We present the results, only for the protocol 1(quantitative criteria), in Fig 1. The legibility has been evaluated by several readers, not involved in the project.

Results (25 texts processed)	**No error**	**1 error**	**2 errors**	**3 errors**	**4 errors**
Dangling anaphora	9 texts	8 texts	2 texts	4 texts	1 text
Irrelevant sentences	19	6	0	0	0
Missing argumentation links	10	7	3	3	2

Results (25 texts processed)	**excellent**	**good**	**poor**
Legibility	6 texts	15 texts	4 texts

Figure 1

7 - Annexe

Here is an extract from the text «LOOSES R., Le charbon : L'énergie d'avenir dans le monde, *Revue de l'énergie*, n° Juin 1995.» (1785 words in the original text and 353 words in the extract). The sign «[...]» shows that sentences le before and after are not contiguous in the original text.

Le charbon : l'énergie d'avenir dans le monde

Les prédictions et prévisions ne se vérifiant que rarement, prédire avenir est un exercice difficile et l'on se doit d'être prudent lorsqu'on se penche sur la courbe d'évolution de la production mondiale de charbon. Le tassement des dernières années est-il temporaire ou s'agit-il d'un retournement de tendance, c'est à cette question que l'auteur va essayer de répondre en présentant les atouts du charbon. Prédire l'avenir est un exercice difficile, voire téméraire.
[...]
les premiers éléments
[...] J'utiliserai une autre approche pour confirmer cette évolution: l'approche obtenue à partir des prévisions de l'AIE production d'électricité. C'est une approche importante quand on sait que 60 % de la production de charbon concourt à la production d'électricité et que par ailleurs 40 % de l'électricité sont générés à partir de charbon.
[...]
Les atouts du charbon
Quels sont les atouts du charbon qui font que ce scénario de croissance se crédibilise ? En premier lieu, les réserves ; les réserves, et en matière de charbon il s'agit de réserves prouvées et économiquement exploitables, sont considérables : près de 200 ans au rythme actuel d'exploitation.
[...]
En second lieu la capacité de production. Le marché charbonnier est en surcapacité de production depuis quelques années.
[...]
les défis du charbon
[...] Reste bien entendu le débat sur les émissions de CO2 et le changement climatique. Certes le CO2 n'est pas le seul gaz à effet de serre, encore faut-il reconnaître que dans une prise en compte de toute la chaîne énergétique depuis la production jusqu'à la combustion en passant par le transport, le charbon n'est probablement pas plus nocif que d'autres sources d'énergie.
[...]
Conclusion
C'est faire preuve d'un optimisme raisonné que d'affirmer que le charbon jouera sans aucun doute un rôle important dans la satisfaction des besoins énergétiques futurs. Il n'y a pas d'obstacles majeurs au développement du charbon qui se fera surtout dans les pays asiatiques à croissances démographiques et économiques rapides, auxquelles nul ne pourra s'opposer. Finalement le charbon, grâce à ses immenses réserves apparaît comme l'énergie refuge, le recours indispensable pour passer de l'energy mix d'aujourd'hui à celui du milieu du prochain siècle.

8 - Bibliography

ADAM J.M., 1990 , *Éléments de linguistique textuelle*, Mardaga, Liège.

BERRI J., LE ROUX, D., MALRIEU D., MINEL, J.L., 1995 , « SERAPHIN un système d'extraction automatique d'énoncés importants », *Actes du colloque Génie linguistique*, Montpellier, pp409-419.

BERRI J., LE ROUX, D., MALRIEU D., MINEL, J.L., 1995, « SERAPHIN main sentences automatic extraction system », *Second Language Engineering Convention* . Londres

BERRI, J., CARTIER E., DESCLES J-P, JACKIEWICZ A., MINEL J.L., 1996, « Filtrage Automatique de textes ». *Natural Language Processing and Industrial Applications*, pp 28-35, Moncton, N-B, Canada.

BERRI J., 1996, « Contribution à la méthode d'exploration contextuelle, applications au résumé automatique et aux représentations temporelles; réalisation informatique du système SERAPHIN», *Thèse de l'Université Paris Sorbonne* . Paris.

BOURIGAULT D., 1994, « LEXTER, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes», *Thèse de l'EHESS*. Paris

CARTIER E., 1997, *LA DÉFINITION : ses formes d'expression, son contenu et sa valeur dans les textes*, thèse en cours. Université de Paris Sorbonne. Paris

CHAROLLES M., 1988, «Les plans d'organisation textuelle , période, chaînes, portées et séquences», *Pratiques*, n 57, Metz.

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.

CHOUEKA T., LUSIGNAN S., 1985, « Disambiguation by short contexts », *Computer and Humanities*, 19, 3, pp 147-157

DE JONG G., 1982, « An overview of the FRUMP system », in , *Strategies for Natural Language Processing*/ W.G. Lehnert & M.H. Ringle [eds], London , Lawrence Erlbaum, pp. 149-172.

DESCLES J-P, JOUIS C, MAIRE-REPPERT D, OH H-G, 1991, « Exploration contextuelle et sémantique , Un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte », in *Knowledge modeling and expertise transfert*, D. Hérin-Aime, R. Dieng, J.P. Regourd, J.P. Angoujard [éds], 371-400, Amsterdam, Washington DC, Tokyo, IOS Press.

DESCLES J.P., BERRI J., JACKIEWICZ A., MALRIEU D., MINEL, J-L., 1995, « Le résumé automatique par exploration contextuelle », Rapport CAMS, 62 p.

ENDRES-NIGGEMEYER, B., 1996, *Summarising text*, à paraître.

FUCHS C, DANLOS L., LACHERET-DUJOUR A., LUZZATI D., VICTORRI B., 1993, *Linguistique et traitement automatique des langues*, Hachette , Paris.

HAHN U., REIMER U., 1985, The TOPIC project , text-oriented procedures for information management and condensation of expository texts, University of Constance.

JACKIEWICZ A., 1996 , « La notion de cause pour le filtrage de phrases importantes d'un texte», *Natural Language Processing and Industrial Applications*, pp136-141, Moncton, N-B, Canada.

JACKIEWICZ A., 1997, « Modélisation des connaissances extraites des documents techniques. Le poblème de la causalité.» Thèse en cours. Université de Paris Sorbonne. Paris

KÄLLGREN G., 1988, « Automatic abstracting of content in text », *Nordic Journal in Linguistics*, 11, pp. 89-110.

LEHNERT W.G., 1981, « Plot units and narrative summarization », *Cognitive Science*, 5, pp. 293-331.

LE ROUX D., 1991, « Automatisation de l'activité résumante , essai de typologie », *Actes du colloque international sur le résumé de texte*, Pont-à-Mousson / Université de Nancy II, septembre 1991, Klincksieck Ed.

LE ROUX D., MINEL J.L., BERRI J., 1994, « SERAPHIN Project , the industrial approach », *Actes du colloque «Cognitive Science in industry»*, pp 275-283.

MARANDIN, J.M., 1993, «Analyseurs syntaxiques. Équivoques et problèmes » *TAL Analyse syntaxique* 34, 1, pp. 5-33.

MIIKE S, ITOH, E. & al., 1994, « A full-text retrieval system with a dynamic abstract generation function », *SIGIR 94*, Dublin, pp. 152-161 .

PAICE C.D., 1990, « Constructing literature abstracts by computer , techniques and prospects », *Information processing management*, 26 [1], pp. 171-186.

PERELMAN C, OLBRECHTS-TYTECA L, 1992, *Traité de l'argumentation*, Editions de Bruxelles.

PLANTIN C., 1990, *Essais sur l'argumentation*, Editions Kimé, Paris.

PUGEAULT F, 1995, « Extraction dans les textes de connaissances structurées: une méthode fondée sur la sémantique lexicale », Ph. D., IRIT Université Paul Sabatier, Toulouse 3, n° 2153.

RENOUF A, J., COLLIER A., 1995, « A system of automatic textual abridgement », *Actes du colloque «Génie linguistique«*, Montpellier, pp 395-407

ROULET E. et alii, 1985, *L'articulation du discours en français contemporain*, Bern, Peter Lang.

ROULET E., 1987, « Complétude interactive et connecteurs reformulatifs », *Cahiers de linguistique française*, n°8, pp.111-140.

SABAH G., 1988, L'*intelligence artificielle et le langage, représentation des connaissances*, HermèsC, Paris

SAINT-DIZIER P., 1995, « Constraint propagation techniques for lexical semantics descriptions » in *Computational semantics*, Saint-Dizier P., Viegas E., Cambridge University New York pp 426-440.

SALTON, G.; 1989; Automatic text processing : the transformation, analysis and retrieval of information by computer; Addison Weshley Publ. Comp.

SLATOR, B., 1991, «Using context for some preference », in *Lexical Acquisition , Exploiting On-Line Resources to build a Lexicon*, Zernik, U. Ed., Lawrence Earlbaum, Hillsdale, NJ