Université de Montréal

Génération automatique de résumés par analyse sélective

par

Horacio Saggion

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l’obtention du grade de
Philosophiae Doctor (Ph.D.)
en Informatique

Août 2000

©Horacio Saggion, 2000
Université de Montréal
Faculté des études supérieures

Cette thèse intitulée

Génération automatique de résumés par analyse sélective
présentée par

Horacio Saggion

a été évaluée par un jury composé des personnes suivantes :

Jian-Yun Nie
.................................
(président-rapporteur)

Guy Lapalme
.................................
(directeur de recherche)

Michel Boyer
.................................
(membre du jury)

Eduard Hovy
.................................
(examineur externe)

Thèse acceptée le :

.................................
Sommaire

Un résumé est un texte concis qui rend compte du contenu “essentiel” d’un document. Son objectif est d’aider le lecteur à cerner la pertinence du document par rapport à l’information recherchée. Bien que l’idée de produire des résumés de manière automatique ne soit pas nouvelle, de nos jours, avec le volume croissant de textes électroniques disponibles, le développement d’outils informatiques pour la production de résumés de textes se fait sentir.

Mais comment faire pour qu’un ordinateur puisse calculer le contenu essentiel d’un document et l’exprimer sous la forme d’un nouveau texte cohérent et cohérent? C’est la problématique du résumé automatique abordée dans cette thèse. Produire des résumés est une tâche fort difficile, car elle nécessite des connaissances linguistiques et du monde qui ne sont pas faciles à incorporer dans des systèmes automatiques.

En explorant cette problématique, cette thèse propose une nouvelle méthode de production de résumés pour le texte technique. Notre démarche méthodologique a consisté à étudier un corpus de résumés techniques et leurs documents sources. L’objectif de cette étude fut de répondre à deux questions fondamentales : comment sélectionner le contenu du résumé et comment l’exprimer. Nous montrons quelles parties structurales du document source sont le plus souvent utilisées pour repérer le contenu essentiel ; quel est le statut de ce contenu et comment il est reformulé sous la forme d’un nouveau texte. Nous avons défini un modèle linguistique et conceptuel qui identifie des concepts et des relations avec des marqueurs linguistiques. Notre étude fait donc ressortir les liens entre le texte source et le résumé dans le cas particulier des textes techniques et scientifiques.


**SumUM**, notre prototype, a été développé pour démontrer la viabilité de notre approche. Nous y avons implanté quelques fonctionnalités de l’Analyse Sélective avec des techniques robustes de traitement du langage naturel tels que des programmes d’étiquetage lexi-
cal, l’application d’automates à états finis pour reconnaître des constructions syntaxiques simples, la classification des phrases, l’extraction d’informations, l’instanciation de patrons et la régénération de textes. Ces techniques nous ont aidé à implanter quelques phénomènes observés lors de notre étude.

Nous avons démontré la viabilité de notre approche au moyen de trois évaluations au cours desquelles nous avons comparé les résumés produits par SumUM avec d’autres technologies de production de résumés. À l’aide d’évaluations avec des sujets humains, nous avons pu montrer que la qualité des textes produits par l’Analyse Sélective est supérieure à celle d’autres systèmes d’extraction de phrases.

**Mot clés :** traitement du langage naturel, résumé automatique, régénération de textes, évaluation de résumés
Abstract

An abstract is a text of a recognizable genre with a very specific purpose: to give the reader an exact and concise knowledge of the contents of a document. Nowadays, the overwhelming quantity of information and the need to access the essential content of documents accurately to satisfy users' demands calls for the development of computer programs able to produce text summaries. This is a difficult task because it requires linguistic and world knowledge which are difficult to incorporate in automatic systems.

In this dissertation, we explore the problem of automatic text summarization and propose a new method of text summarization of technical articles. Our research method consisted of the study of a corpus of abstracts written by professional abstractors. We have studied relations between abstracts and their source documents in order to answer the following questions: where does the information reported in abstracts come from? how can it eventually be found in source documents? what is its status? and how is it conveyed in the abstract? We show that the information brought into the technical abstract has a well-defined status and can be identified on the basis of general concepts and relations. Based on this study, we have defined a conceptual and linguistic model for the task of technical text summarization.

Selective Analysis, our method for text summarization, produces abstracts relying on the structure of the source document and on specific types of information from the conceptual model. The method was designed to produce short summaries in two steps: first, the reader is presented with an abstract which identifies the topics of the document. Then, if the reader is interested in some of the topics, additional information about them is presented. The abstracts are produced by the processes of conceptual identification and text regeneration.

SumUM, our prototype, was developed in order to validate this methodology of text summarization. We have implemented some functionalities of Selective Analysis using state of the art techniques in natural language processing such as text segmentation, part of speech tagging, partial syntactic and semantic analysis, template instantiation and text regeneration.

We have evaluated Selective Analysis by comparing the abstracts produced by SumUM with abstracts produced using other summarization methodologies. In an evaluation of content, we show that the automatic abstracts help readers in a classification task and also that SumUM selects sentences considered more important to readers than those selected by other summarization technologies. An evaluation of text quality with human informants
demonstrates that the abstracts produced with SumUM are of better quality than abstracts produced by a sentence extraction methodology.

Keywords: Natural Language Processing, Automatic Abstracting, Text Re-generation, Evaluation
# Table des matières

1 **Introduction** 1  
1.1 Résumons 2  
1.2 Plan de la thèse 4  

2 **Les résumés** 5  
2.1 Types de résumés 5  
2.2 Organisation 7  
2.3 Production des résumés 9  
2.3.1 Macro Structures 9  
2.3.2 History Grammars 11  
2.3.3 Plot Units 11  
2.3.4 Concept Coherence 12  
2.3.5 Production chez les rédacteurs professionnels 13  
2.4 Conclusion 15  

3 **Le résumé automatique** 17  
3.1 Introduction 17  
3.2 Les méthodes d’extraction de phrases 18  
3.2.1 Méthode de distribution de termes 18  
3.2.2 La méthode de la position 21  
3.2.3 Expressions indicatives 21  
3.2.4 Analyse rhétorique 23  
3.2.5 Cohésion lexicale 28  
3.2.6 Classification des éléments 29  
3.3 Approches hybrides 31  
3.3.1 Combinaison manuelle 31  
3.3.2 Combinaison statistique 31  
3.4 Le problème de la cohésion 34  
3.5 Méthodes de compréhension et génération 34  
3.5.1 Scénarios 35  
3.5.2 Instanciation de patrons et génération 37  
3.6 Conclusion 39  

vii
# 4 Observations from the Corpus

4.1 Human Produced Abstracts .............................................. 41

4.2 Where is the Topic? .................................................................. 42

4.2.1 Corpus Description .......................................................... 42

4.2.2 Corpus Analysis ............................................................... 42

4.2.3 Tables of Alignment ........................................................ 44

4.2.4 Distributional Results ...................................................... 49

4.2.5 Analysis of the Results ..................................................... 50

4.3 Conceptual and Linguistic Model ............................................ 52

4.4 From Source to Abstract ...................................................... 55

4.5 Summary ............................................................................ 63

# 5 Selective Analysis

5.1 Introduction ........................................................................ 65

5.2 The Input ........................................................................... 67

5.3 Pre-processing and Interpretation ......................................... 68

5.4 Representing the Information for the Abstract ...................... 72

5.4.1 Indicative Templates ....................................................... 72

5.4.2 Informative Templates ..................................................... 86

5.5 Indicative Selection ............................................................. 89

5.6 Informative Selection .......................................................... 91

5.7 Generation ........................................................................ 92

5.8 A short annotated example . ....... ....................................... 102

5.9 A longer example .............................................................. 105

5.10 Discussion ....................................................................... 109

5.11 Summary ....................................................................... 112

# 6 Implementing Selective Analysis in SumUM

6.1 Introduction ....................................................................... 113

6.2 Pre-Processing ................................................................... 114

6.3 Interpretation ..................................................................... 116

6.4 Indicative Selection ............................................................ 124

6.4.1 Extraction using Patterns ............................................... 125

6.4.2 Extraction using Domain Relations .................................. 128

6.4.3 Selecting the Content .................................................... 131

6.5 Informative Selection .......................................................... 131

6.6 Generation ....................................................................... 132

6.7 Limitations of the Implementation ....................................... 136

6.8 Summary ....................................................................... 138

# 7 Evaluating Content and Text Quality in Selective Analysis

7.1 Introduction ....................................................................... 139

7.2 First Evaluation : Human Abstracts as Ideal Abstracts ........... 140

7.2.1 Experiment 1 ................................................................. 141

7.2.2 Results of Indicativeness ............................................... 145


7.2.3 Influence of the Term Extraction Algorithm in the Evaluation . . . . 147
7.2.4 Content-Based Measures of Evaluation .......................... 148
7.2.5 Acceptability ......................................................... 148
7.2.6 Experiment 2 ........................................................ 149
7.2.7 Result of Acceptability ............................................. 149
7.3 Second Evaluation: Extrinsic ........................................... 150
4.3.1 Evaluation at Université de Montréal ................................ 151
7.3.2 Evaluation at McGill University ..................................... 155
7.3.3 Evaluation at John Abbott College ................................. 157
7.4 Third Evaluation: Informativeness ..................................... 159
7.4.1 Materials ................................................................ 160
7.4.2 Procedure .................................................................. 161
7.4.3 Results of Summarization Systems ................................. 162
7.4.4 Comparison of Human Summaries ................................. 163
7.5 Discussion ................................................................. 164
7.6 Summary ................................................................. 166

8 Conclusion et Perspectives ................................................ 167

Bibliographie ................................................................. 172

A Journaux source utilisés pour le corpus .................................. 183

B Expressions indicatives identifiées dans le corpus ...................... 185

C Concepts ....................................................................... 189

D Relations ....................................................................... 193

E Types of Information in Selective Analysis ............................... 197
  E.1 Indicative Types .......................................................... 197
  E.2 Informative Types ........................................................ 202

F Parsed Segment Fragments ................................................. 205

G Instruction Booklet for the Evaluators .................................. 209
  G.1 Overview .................................................................. 209
  G.2 Evaluation ............................................................... 209

H Informed Consent Form to Participate in Research ................... 211

I Questionnaire Sample for Evaluation of Indicativeness ................ 213

J Abstracts Produced by SumUM .......................................... 221
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Margie : texte court de type narratif et son résumé</td>
<td>2</td>
</tr>
<tr>
<td>2.2</td>
<td>Digest du service CBC Newsworld Online News Digest pour l’article “HEAVY TURNOUT IN ZIMBABWE ELECTION”, June 25, 2000</td>
<td>6</td>
</tr>
<tr>
<td>2.4</td>
<td>Synopsis du film “Psycho” d’Alfred Hitchcock</td>
<td>6</td>
</tr>
<tr>
<td>2.7</td>
<td>Résumé PASCAL numéro 5802 de l’article “Analyse exacte et en moyenne d’algorithms de recherche d’un motif dans un texte.”</td>
<td>8</td>
</tr>
<tr>
<td>2.8</td>
<td>Résumé de l’article “Angiotensin II modulates conducted vasoconstriction to norepinephrine and local electrical stimulation in rat mesenteric arterioles”, Gustafsson, F., Cardiovascular Research, Vol 44, Issue 1, October 1999, 176-184</td>
<td>9</td>
</tr>
<tr>
<td>2.9</td>
<td>Interprétation de Margie selon les History Grammars</td>
<td>12</td>
</tr>
<tr>
<td>2.10</td>
<td>Interprétation de Margie selon les Concept/Coherence</td>
<td>13</td>
</tr>
<tr>
<td>2.11</td>
<td>Résumé indicatif 63034 de la revue Computer &amp; Control Abstracts pour les annales du CSL’89 3rd Workshop on Computer Science Logic</td>
<td>14</td>
</tr>
<tr>
<td>3.1</td>
<td>Schemas RST</td>
<td>24</td>
</tr>
<tr>
<td>3.2</td>
<td>Arbres RST</td>
<td>25</td>
</tr>
<tr>
<td>3.3</td>
<td>Attribution de poids dans Ono</td>
<td>26</td>
</tr>
<tr>
<td>3.4</td>
<td>Promotion des nœuds dans Marcu</td>
<td>26</td>
</tr>
<tr>
<td>5.1</td>
<td>Main Processes in Selective Analysis</td>
<td>66</td>
</tr>
<tr>
<td>5.2</td>
<td>Sample Input Text</td>
<td>68</td>
</tr>
<tr>
<td>5.3</td>
<td>Specification of the Situation Template</td>
<td>73</td>
</tr>
</tbody>
</table>
5.4 Specification of the Problem Identification Template ..................... 74
5.5 Specification of the Problem Solution Template .......................... 74
5.6 Specification of the Need for Research Template ......................... 75
5.7 Specification of the Entity Introduction Template ....................... 75
5.8 Specification of the Topic of Document Template ........................ 76
5.9 Specification of the Topic Description Template .......................... 76
5.10 Specification of the Possible Topic Template ............................. 77
5.11 Specification of the Research Goal Template ............................. 77
5.12 Specification of the Conceptual Goal Template ............................ 78
5.13 Specification of the Concept Template ..................................... 78
5.14 Specification of the Conceptual Focus Template .......................... 79
5.15 Specification of the Author Development Template ........................ 79
5.16 Specification of Development Template ..................................... 80
5.17 Specification of the Author Interest Template ............................. 80
5.18 Specification of the Author Study Template ................................ 81
5.19 Specification of the Study Template ........................................ 81
5.20 Specification of the Method Template .......................... 82
5.21 Specification of the Experiment Template ................................ 82
5.22 Specification of the Result Template ...................................... 83
5.23 Specification of the Inference Template .................................. 83
5.24 Specification of the Knowledge Template .................................. 84
5.25 Specification of Summarization Template .................................. 84
5.26 Specification of Entity Identification Template ............................ 84
5.27 Specification of the Topic of Section Template ........................... 85
5.28 Specification of Signaling Structural Template ............................ 85
5.29 Specification of Signaling Concept Template .............................. 86
5.30 Specification of Multiple and Merged Templates .......................... 86
5.31 Specification of Informative Templates .................................... 87
5.32 Instantiation of the Topic of Section Templates and Merging .......... 93
5.33 Indicative abstract and list of topics produced by SumUM for the source document “Features 3D scanning systems for rapid prototyping” from the Journal Assembly Automation, 17(3), 1997 ......................... 102
5.34 Topic elaboration for the abstract in Figure 5.33 ........................ 103
5.35 Instantiated Possible Topic Template ...................................... 105
5.36 Instantiated Definition Template ............................................ 105
5.37 Indicative abstract and list of topics produced by SumUM for the source document “DRAMA, a connectionist architecture for online learning and control of autonomous robots : experiments on learning of a synthetic proto-language with a doll robot” from the Journal Industrial Robots 26(1), 1999 .......... 106
5.38 Informative abstract produced by SumUM for the source document “DRAMA, a connectionist architecture for online learning and control of autonomous robots : experiments on learning of a synthetic proto-language with a doll robot" from the Journal Industrial Robots 26(1), 1999 .......... 107
5.39 Source Document for the Abstract in Figure 5.37 ........................ 108
6.1 Pre-processing and Interpretation in **SumUM** .......................... 115
6.2 Input Text and POS-Tagging .................................................... 115
6.3 Sentence Representation from the Tagged Files .......................... 119
6.4 Syntactic and Conceptual Information in Parsed Sentences .......... 123
6.5 Attribute Pair Values in Parsed Sentence Fragments ................. 124
6.7 Indicative Selection .............................................................. 126
6.8 Specification of Indicative and Informative Sentence Patterns and Prolog Representation of the **Author’s Goal** and **Definition** .......................................................... 127
6.9 **Author’s Goal** Instantiated during Indicative Selection .......... 129
6.10 **Topic of the Document** Instantiated during Indicative Selection .... 130
6.11 Information on the Term Tree .................................................. 132
6.12 Informative Selection ............................................................ 133
6.13 Output of **SumUM** for the article “Climbing, walking and intervention robots”, Industrial Robot, Vol 24 Issue 2, 1997 .......... 134

7.1 Abstracts used for evaluation purposes for the Document “Design and implementation of an aided fruit-harvesting robot (Agribot) ”, Industrial Robot, Vol 25 Issue 5, 1999 (continues on Figure 7.2) ...................... 143
7.3 **SumUM** abstract for the document “Telexistence and R-Cubed”, Industrial Robots 26(3) used for evaluation purposes .................. 151
7.4 **Microsoft’97 Summarizer** abstract for the Document “Telexistence and R-Cubed”, Industrial Robots 26(3) used for evaluation purposes .......... 152
7.5 Abstracts published with the source document “Telexistence and R-Cubed”, Industrial Robots 26(3) used for evaluation purposes .......... 152
<table>
<thead>
<tr>
<th></th>
<th>Résumé des étapes de la méthode de distribution de termes</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>Résumé des étapes de la méthode de la position</td>
<td>22</td>
</tr>
<tr>
<td>3.3</td>
<td>Résumé des étapes de la méthode des expressions indicatives</td>
<td>23</td>
</tr>
<tr>
<td>3.4</td>
<td>Résumé des étapes de la méthode rhétorique</td>
<td>28</td>
</tr>
<tr>
<td>3.5</td>
<td>Étapes de la méthode de cohésion lexicale</td>
<td>29</td>
</tr>
<tr>
<td>3.6</td>
<td>Étapes de la méthode des chaînes lexicales</td>
<td>29</td>
</tr>
<tr>
<td>3.7</td>
<td>Étapes de la méthode d'extraction de paragraphes</td>
<td>29</td>
</tr>
<tr>
<td>3.8</td>
<td>Résumé des étapes de la méthode de classification sémantique</td>
<td>30</td>
</tr>
<tr>
<td>3.9</td>
<td>Résumé des étapes de la méthode probabiliste</td>
<td>34</td>
</tr>
<tr>
<td>3.10</td>
<td>Résumé des étapes de la méthode des scénarios</td>
<td>37</td>
</tr>
<tr>
<td>3.11</td>
<td>Résumé des étapes de la méthode des patrons</td>
<td>39</td>
</tr>
</tbody>
</table>

4.1 Alignment of the professional abstract LISA 3024 and the source document "Movement characteristics using a mouse with tactile and force feedback" International Journal of Human-Computer Studies, 45(5), Oct’96 p483-93 | 44 |

4.2 Alignment of the professional abstract LISA 4600 with the source document "Fuzzy and neural hybrid expert systems: synergic AI", M. Funabashi and others. IEEE Expert, 10(4) Aug 95, p32-42 | 45 |

4.3 Alignment of the professional abstract LISA 6863 with the source document "The DECIMAL project: decision-making and decision support in small to medium size libraries" T. Oulton and others. Vine, (103) 1996, p13-19. (continues on Table 4.4) | 46 |

4.4 (Continues from Table 4.3) Alignment of the professional abstract LISA 6863 with the source document "The DECIMAL project: decision-making and decision support in small to medium size libraries" T. Oulton and others. Vine, (103) 1996, p13-19 | 47 |

4.5 Alignments of Different Sentences from the Corpus (continues on Table 4.6) | 48 |

4.6 (Continues from Table 4.5) Alignments of Different Sentences from the Corpus | 49 |

4.7 Distribution of Information | 49 |

4.8 Concepts in Source Documents and Professional Abstracts | 52 |

4.9 Some Concepts from the Conceptual Model | 53 |

4.10 Some Relations from the Conceptual Model | 54 |

4.11 Text Editing in Human Abstracting | 56 |

4.12 Syntactic Verb Transformation | 57 |

4.13 Lexical Verb Transformation | 57 |
7.5 Sentences from the 3 Sources: **SumUM** (SA), Professional Abstractor (PA), and Source Document (SD) ........................................ 149
7.6 Number of Acceptable Sentences and Average Acceptability .................. 150
7.7 Keywords from Industrial Robots used in the Evaluation at Université de Montréal ......................................................... 153
7.8 Results of Human Judgment about Indicativeness and Text Quality. Data from the evaluation carried out at École de Bibliothéconomie et des Sciences de l'Information de l'Université de Montréal ........................................ 154
7.9 Keywords from Industrial Robots used on the Evaluation at McGill University 156
7.10 Results of Human Judgment about Indicativeness and Text Quality. Data from the evaluation carried out at McGill Graduate School of Library & Information Studies ......................................................... 156
7.11 Results of Human Judgment about Indicativeness and Text Quality. Data from the evaluation carried out at John Abbott College .................. 159
7.12 Comparison between sentences selected by human informants and sentences selected by three automatic summarization methods: general scenario. The columns contain the information about Recall, Precision and F-score ......................................... 162
7.13 Comparison between sentences selected by human informants and sentences selected by three automatic summarization methods: union, intersection, optimistic and pessimistic scenarios. The columns contain the information about average Recall, Precision and F-score ........................................... 163
7.14 Comparison between sentences selected by the first group and sentences selected by the second group. Each row represents a document. The last row is the average information. The columns contain the information about Recall, Precision and F-score ................................................... 164
xviii
Remerciements

D’abord, je tiens à remercier mon directeur de recherche, Guy Lapalme. Il m’a toujours encouragé dans ma démarche scientifique et m’a offert la meilleure ambiance pour mon développement intellectuel. À plusieurs reprises, il a encouragé et fait possible ma participation lors de réunions scientifiques afin de faire connaître mon travail de recherche.

Merci à Jian-Yun Nie, Michel Boyer et Eduard Hovy, les membres du jury, pour avoir accepté d’évaluer cette thèse.

I would like specially to thank Eduard Hovy for his valuable comments and suggestions that help me improve and clarify the present work.

J’aimerais bien remercier tous les collègues du Laboratoire RALI-Incognito pour une ambiance de travail et discussion agréable et décontracté. Un merci spécial à Narjès, Marco, Nikolay, Wessel, Leila, Felipe, Mohamed, Jean-Marc, Graham, George, Michel, Elliott, Stéphane et Christophe. Merci à tous les collègues qui ont participé lors de mes évaluations.

Ma gratitude à Laurence Danlos qui m’a accueilli dans le laboratoire TALANA de l’Université de Paris VII dans le cadre du projet franco-québécois GRABIG.

J’aimerais aussi remercier le personnel de la Bibliothèque de Mathématiques et Informatique pour leurs services toujours efficace et leurs gentillesses et le personnel administratif du Département d’informatique et de recherche opérationnelle. Un gros merci au personnel du Bureau des services aux étudiants étrangers, spécialement à Mme Caroline Reid.

Je remercie Mme Michèle Hudon et Mme Gracia Pagola de l’EBSI Université de Montréal, professeur John E. Leide de l’Université McGill et Mme Christine Jacobs du Collège John Abbott pour leur intérêt à mon projet de recherche. Je tiens à remercier tous ceux qui ont évalué les résumés produits par mon système.

Je tiens à remercier les institutions suivantes pour leur soutien financière lors de différentes étapes de mon doctorat : Association Canadienne de Développement Internationale, Université de Montréal, Ministerio de Educación de la Nación de la República Argentina, Departamento de Computación de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires et Fundación Antorchas.
Gostaria de agradecer à turma do Brasil et “do mundo” pelos bons momentos que passamos juntos nesta maravilhosa Montréal. Um obrigado muito especial para Anderson e Eliany pela sua amizade nestes anos.

A mis queridos Mamá, Clarisa y Fabio por el apoyo constante durante esta ausencia.

Gracias, obrigado, merci, thanks Sandra por acompañarme con tu amor en una aventura mas.
A vos Sandra y a tu precioso cargamento
Chapitre 1

Introduction

Un résumé est un texte concis qui rend compte du contenu “essentiel” d’un autre texte, dit texte source. Son objectif est d’aider le lecteur à cerner la pertinence du document source vis-à-vis de l’information recherchée. Après la lecture d’un résumé, le lecteur sera en position de décider si le document contient l’information pertinente pour sa tâche actuelle ou pas. Il se peut aussi que le lecteur n’ait pas besoin de lire le document original simplement parce que les informations recherchées sont comprises dans le résumé.

Dans la plupart des cas, les résumés sont produits par des humains mais, de nos jours, la quantité de texte disponible en format électronique est énorme (en 1998 on a calculé son volume quelque part entre 400 et 500 millions de documents (Filman and Pant, 1998)) et en conséquence l’incorporation d’outils informatiques pour la production de résumés de textes se fait sentir.

En particulier, au cours de la dernière décennie plusieurs événements scientifiques se sont plongés dans la problématique de la production automatique de résumés (Dagstuhl, 1993; ACL/EACL, 1997; AAAI, 1998; RIFRA, 1998; WAS, 2000) et la communauté scientifique réunie lors du Dagstuhl Seminar a identifié des axes de recherche nécessaires dans ce domaine. Parmi les questions soulevées se trouvent l’étude des différents types de résumés et de leurs fonctions, l’étude de la relation entre texte source et résumé, le développement de procédures d’évaluation du résumé automatique et l’étude des stratégies mixtes (statistiques et symboliques) dans la résolution de ce problème.

Tout en explorant quelques aspects de la problématique du résumé automatique, cette thèse propose une nouvelle méthode de production de résumés pour le texte technique. Bien que l’idée de produire des résumés de manière automatique ne soit pas nouvelle, les résumés produits à l’heure actuelle par ordinateur n’ont pas la qualité des résumés rédigés par un humain. Dans l’état actuel de l’art, les résumés automatiques sont le plus souvent, des extraits du texte source ; les techniques et méthodes utilisées ne garantissent pas la conservation du contenu du texte source et ceci peut entraîner le manque de cohésion, voire de cohérence. Mais ceci a une raison d’être, la tâche de produire des bons résumés est très difficile car elle nécessite la compréhension du texte source qui ne peut pas être atteint avec les techniques
actuelles de traitement du langage naturel, sauf pour des domaines limités. Les limitations actuelles ne peuvent pas être ignorées lors de la conception des systèmes de production de résumés et ceci a grandement influencé notre travail de recherche.

1.1 Résumons

Un résumé est obtenu à partir des opérations de :
- sélection de ce qui est important dans le texte source,
- condensation de cette information par généralisation et élimination des répétitions et,
- expression du contenu propositionnel qui reste sous la forme d’un texte cohésif.

Dans ce processus il y a, évidemment, une perte d’information mais le message original et sa cohérence sont préservés. Dans la Fig. 1.1 on présente le texte narratif “Margie” et son résumé1.

Margie serrait son beau et nouveau ballon par la corde. Tout à coup, une rafale le captura. Le vent l’entraîna vers un arbre. Le ballon frappa une branche et explosa. Margie pleura et pleura.

*Margie pleura quand le vent cassa son ballon.*

Fig. 1.1: Margie : texte court de type narratif et son résumé

On peut voir que le résumé préserve le contenu essentiel de l’histoire. Les informations retenues (i.e. sélectionnées) à inclure dans le résumé sont : l’état final de la protagoniste : elle pleure, et les causes de cet état : son ballon a crié. Il n’y a pas d’information circonstancielle : ni le fait que le ballon était nouveau et beau, ni le fait que Margie serrait le ballon avant qu’il crève, ni le fait que l’instrument causant le déchirement du ballon était une branche d’arbre ne sont inclus dans le résumé.

La production d’un résumé dépend de l’interprétation du texte source : reconnaissance des relations entre les informations exprimées dans le texte source et déduction, entre autres. Il n’y a pas un seul résumé possible d’un texte car il dépend entre autres de l’état de connaissances de l’interprète, de ses objectifs, et de son état d’attention.

Produire des résumés est une faculté cognitive de l’être humain : lors de la lecture d’un texte, certaines propositions sont “retenues” par le lecteur alors que d’autres sont “oubliées” : on produit des résumés de manière naturelle dans le cadre du processus de compréhension. Mais, les résumés sont aussi produits dans des contextes professionnels par des “rédacteurs de résumés.” Ces professionnels sont capables de rédiger des textes très succincts (par rapport au texte source) pour une “audience” bien ciblée. Dans la Fig. 1.2 on présente un résumé

---

1Le texte est une traduction libre en français d’un exemple de Rumelhart (1975).
1.1. RÉSUMONS

d’un texte de type scientifique de neuf pages. Le résumé a été produit par un rédacteur des résumés de la revue Library & Information Science Abstracts. Le texte est très succinct, il présente à peine les thèmes abordés dans l’article. Dans notre travail de recherche nous nous intéressons surtout aux résumés des textes scientifiques et techniques tel que celui présenté à la Fig. 1.2.

Present a more efficient Distributed Breadth-First Search algorithm for an asynchronous communication network. Presents a model and gives an overview of related research. Analyzes the complexity of the algorithm, and gives some examples of performance on typical networks.

Fig. 1.2: Résumé de l’article : Efficient distributed breadth-first search algorithm. S.A.M. Makki. Computer Communications, 19(8) Jul 96, p628-36

Notre objectif est de produire des résumés de manière automatique à partir du texte source. Mais, afin de comprendre ce qu’est un résumé et comment les informations essentielles du texte source peuvent être identifiées et exprimées, nous avons décidé d’étudier des exemples réels de résumés.

Notre démarche méthodologique consiste à étudier un corpus de résumés techniques et leurs documents sources. Les résumés ont été rédigés par des rédacteurs professionnels. Nous avons choisi d’étudier ce type de résumés car, généralement, ils sont mieux structurés que les résumés d’auteurs et aussi parce qu’en général ils respectent les mots de l’auteur. Il est important de noter que dans cette thèse nous ne nous intéressons pas au processus cognitif de la production des résumés par des professionnels mais que nous utilisons les résumés pour comprendre comment on pourrait les produire de manière automatique. Ces processus cognitifs ont déjà été étudiés dans les travaux de Endres-Niggemeyer et al. (1991, 1995).

Par contre, la relation entre texte source et résumé que nous abordons dans cette thèse a été généralement négligée. Nous explorons les aspects suivants de la relation entre document source et résumé : (i) où l’information du résumé se trouve-t-elle dans le document source ? (ii) quel est le statut de cette information ? (iii) comment peut-elle être repérée ? et (iv) comment les informations sont-elles utilisées pour construire le résumé ? L’objectif de cette étude est de répondre à deux questions fondamentales dans la problématique du résumé automatique : comment sélectionner le contenu du résumé et comment l’exprimer. Nous n’adressons cette question que dans le contexte du texte technique et nous le faisons en analysant un nombre limité de données.

Notre étude de corpus indique qu’il y a certains types d’information qui sont généralement retenus pour un résumé et que cette information peut être repérée en utilisant des marqueurs linguistiques et de position dans le texte source. À partir de cette étude, nous avons défini une méthode calculatoire pour la production des résumés qui se base fortement sur la structure du texte et sur des marqueurs linguistiques.
Notre méthode propose à l'utilisateur d'abord un résumé qui introduit certains thèmes considérés importants dans le texte source et ensuite, l'utilisateur peut décider d'explorer ces thèmes en demandant plus d'information du texte source telles que définitions et descriptions. De cette manière, nous abordons la production d'un résumé dynamique en fonction de l'intérêt du lecteur. Le résumé est produit par un processus d'identification conceptuelle et de re-génération.

1.2 Plan de la thèse

La thèse est organisée de la manière suivante. Dans le chapitre 2, nous introduisons d'abord la terminologie nécessaire et le processus de production des résumés chez les humains. Au chapitre 4, nous détaillons quelques approches pour la production de résumés automatiques. Parmi l'éventail des approches existantes, nous n'en avons choisi que quelques-unes qui montrent l'état d'avancement de la recherche et qui ont influencé le développement de notre méthode. Le chapitre 4, détaille l'analyse du corpus et le modèle linguistique et conceptuel développé pour la tâche de production des résumés de type technique. L'Analyse Sélective, notre méthode pour produire des résumés basée sur les résultats du chapitre 4, est introduite au chapitre 5. Un prototype computationnel a été développé pour montrer la viabilité de l'approche. Celui-ci implante quelques fonctionnalités du modèle théorique avec des techniques robustes de traitement du langage naturel telles que la reconnaissance des groupes nominaux et verbaux de base, la fouille de modèles et l'instanciation de patrons, et quelques techniques de re-génération de textes. Le prototype est décrit au chapitre 6. Le chapitre 7 présente le problème de l'évaluation de résumés automatiques, en particulier nous nous intéressons à l'évaluation en utilisant des juges formés en science de l'information. Finalement, des conclusions et un panorama des travaux futurs sont présentés au chapitre 8.

Chapitre 2

Les résumés

Dans ce chapitre nous nous intéressons aux types de résumés et au processus de production de résumés chez les humains. Plus spécifiquement nous introduisons les notions de résumé indicatif et informatif qui constituent le sujet de notre recherche et nous explorerons brièvement les processus de compréhension et condensation tels qu'étudiés en linguistique textuelle, science cognitive, intelligence artificielle et science de l’information.

2.1 Types de résumés

Dans cette recherche nous nous intéressons exclusivement au résumé du texte technique et scientifique qui en langue anglaise est appelé abstract. Selon Rowley (1982) l’abstract est différent d’autres types de substitut du document source tels que l’extrait qui est un ou plusieurs passages tirés d’un texte; le sommaire qui est une énumération des points principaux d’un discours (voir Figure 2.1); l’abrégé qui est un écrit réduit aux points essentiels; le précis qui est un exposé exact et succinct d’un texte (voir par exemple Russell (1988)); la paraphrase qui est une interprétation des idées de l’auteur du document source qui seront exprimées dans le langage de l’auteur de la paraphrase; le digest qui est un condensé d’un livre ou d’un article journalistique (voir Figure 2.2); le highlight qui est un commentaire ajouté dans des parties spécifiques d’un document pour attirer l’attention du lecteur sur les points importants (voir Figure 2.3); et le synopsis qui, plutôt appliqué en cinématographie, est un récit très bref constituant un schéma de scénario (voir Figure 2.4). Les types de résumés ont aussi été étudiés dans (Pouzet, 1981; Hadjadj and Russeau-Payen, 1981).

The specific improvements offered by SIP over IP can be itemized as follows:

- Larger, more hierarchical addresses to support long-term growth of the Internet.
- “Cluster addresses” for more powerful source-directed routing.

[...]
There were complaints of intimidation but few reports of violence during the first day of Zimbabwe's crucial parliamentary elections. Polls will reopen Sunday for a last round of voting before results are tallied.

**Fig. 2.2:** Digest du service CBC Newsworld Online News Digest pour l’article “HEAVY TURNOUT IN ZIMBABWE ELECTION”, June 25, 2000

Neuronet provides immediate access to real-time life-critical data being acquired at multiple sites across the health center and allows one neurophysiologist to simultaneously monitor multiple surgical procedures.

**Fig. 2.3:** Highlight ajouté à l’article “Neuronet : A Distributed Real-Time System for Monitoring Neurophysiologic Function in the Medical Environment”, Krieger, D. et al, Computer, March 1991, 45-55

Alfred Hitchcock’s landmark masterpiece of the macabre stars Antony Perkins as the troubled Norman Bates, whose old dark house and adjoining motel are not the place to spend a quiet evening. No one knows that better than Marion Crane (Janet Leigh), the ill-fated traveler whose journey ends in the notorious “shower scene.” First a private detective, then Marion’s sister (Vera Miles) search for her, the horror and suspense mount to a terrifying climax when the mysterious killer is finally revealed.

**Fig. 2.4:** Synopsis du film “Psycho” d’Alfred Hitchcock

Dans le contexte du texte technique est scientifique, deux types de résumés sont considérés (AFNOR, 1984; ANSI, 1979; ERIC, 1980; Maizell et al., 1971) : les résumés indicatifs et les informatifs. L’objectif d’un résumé indicatif est de signaler au lecteur le type d’information qui apparaît dans le texte source. Un résumé informatif rapporte des informations du texte source comme des résultats ou des conclusions. En général on trouve des résumés qui combinent les caractéristiques de ces deux types. Il faut mentionner qu’en science de l’information, on identifie aussi le résumé critique où le rédacteur a pour tâche de commenter l’information originale et de la mettre en contexte.

The work of Consumer Advice Centres is examined. The information sources used to support this work are reviewed. The recent closure of many CAC’s has seriously affected the availability of consumer information and advice. The contribution that Public libraries can make in enhancing the availability of consumer information and advice both to the public and other agencies involved in consumer information and advice, is discussed.

**Fig. 2.5:** Résumé indicatif de l’article : “Consumer information and advice : the role of public libraries” J. Rowley, D. Butcher and C. Turner, Aslib proceedings 32 (11/12), December 1980, 417-424
2.2. ORGANISATION

Dans la Fig. 2.5 on trouve un résumé indicatif produit par un rédacteur professionnel. On peut effectivement vérifier que l’information dans le texte du résumé est à peine signalée. La phrase “The work of Consumer Advice Centres is examined” n’inforne pas sur le travail des Centres de Conseil au Consommateur, mais signale que l’information est élaborée dans le texte source. Le même phénomène peut être observé dans la deuxième et quatrième phrase du résumé. Par contre la troisième phrase informe sur un problème : on peut apprendre que certains Centres de Conseil au Consommateur ont fermé et que ce fait a une incidence sur la disponibilité d’information et de conseil pour les consommateurs.

Dans la Fig. 2.6 on présente un résumé informatif du même article. Le résumé contient de l’information sur les tâches accomplis par les CACs, sur les sources d’information utilisées ainsi que sur la contribution des bibliothèques pour résoudre les problèmes soulevés.

An examination of the work of Consumer Advice Centres and of the information sources and support activities that public libraries can offer. CACs have dealt with pre-shopping advice, education on consumers’ rights and complaints about goods and services, advising the client and often obtaining expert assessments. They have drawn on a wide range of information sources including case records, trade literature, contact files and external links. The recent closure of many CAC’s has seriously affected the availability of consumers information and advice. Public libraries can make many kinds of information sources more widely available, both to the public and to the agencies now supplying consumer information and advice. Libraries can cooperate closely with advice agencies through local coordinating committee, shared premises, joint publicity referral and the sharing of professional expertise.


Dans cette thèse nous nous intéressons surtout aux résumés indicatifs et informatifs que nous étudions en détails au chapitre 4. La différenciation entre les types informatif et indicatif semble être applicable dans le contexte de textes informatifs. Il serait rare, bien que possible, de trouver un résumé comme “On décrit l’état d’une fille et les causes.” pour l’histoire de Margie présenté à la Fig. 1.1. Il y a donc des contraintes qui empêchent la génération d’un résumé si bizarre.

2.2 Organisation

Dans la mesure où les résumés sont un genre textuel, l’information et l’organisation de l’information dans un résumé de texte scientifique est prévisible (Bhatia, 1993). En général, un résumé d’un article scientifique contient les catégories d’informations suivantes (Rowley, 1982) :
- Objectif : description de la problématique, les objectifs de l'auteur, les limites du travail, ce que l'auteur a étudié, etc.

- Méthodologie : description de la méthode employée, les variables, les sujets, les procédures, etc.

- Résultats : les valeurs mesurées, etc.

- Conclusions : les relations entre les variables, conséquences pour la théorie, les nouveaux problèmes, etc.

Dans la Fig. 2.7 on montre un résumé d'un article de science et technique publié par la revue PASCAL Bibliographie Internationale. Les catégories d'information incluses dans le résumé sont : l'objectif de l'article dans les phrases (1) à (3) : le problème étudié et quelques particularités ; la méthodologie employée pour obtenir les résultats dans la phrase (4) : les chaînes de Markov sont utilisées, et les résultats obtenus dans les phrases (5) et (6) : la mesure de complexité est donnée.

(1) Nous étudions le problème de la recherche d'un mot, appelé le motif, dans un autre, appelé le texte. (2) Nous considérons l'analyse exacte et l'analyse en moyenne des algorithmes. (3) En particulier, nous donnons les bornes optimales lorsque seulement une tête de lecture se déplaçant de la gauche vers la droite du texte est utilisée : (2-1/m)n comparaisons entre caractères pour tout le texte, et min (1 + log m, card(A)) comparaisons sur chaque caractère de texte, avec m la longueur du motif, n celle du texte, log le logarithme binaire, et A un alphabet général peut être inconnu des algorithmes. (4) Nous utilisons les chaînes de Markov pour l'analyse en moyenne. (5) Nous prouvons que la complexité de la plupart des algorithmes peut être exprimée sous la forme Kn + o(n), où K ne dépend pas de n. (6) Des analyses précises de quelques algorithmes sont aussi données.

Fig. 2.7: Résumé PASCAL numéro 5802 de l'article “Analyse exacte et en moyenne d'algorithmes de recherche d'un motif dans un texte.”

L'information contenue dans un texte scientifique est parfois tellement prévisible que des résumés structurés (Hartley et al., 1996) dans lesquels des étiquettes sont ajoutées au texte du résumé (voir Figure 2.8) ont été proposés pour faciliter le répérage de l'information.
Objective: Localized application of a vasoconstricting agent onto the wall of an arteriole results not only in [...] We investigated the effect of intravenous infusion of angiotensin II (ANG II) [...]

Methods: In anesthetized male Wistar rats (n=43) NE (0.1 mM) or a local depolarizing current was continuously applied onto mesenteric arterioles using micropipettes [...]

Results: Systemic infusion of ANG II (4 ng/min) raised mean arterial blood pressure by 6±2 mm Hg [...] Conclusion: The findings suggest that conducted vasoconstriction to NE and local electrical stimulation in rat mesenteric [...]
une macro-structure composée formellement d’un épisode, qui contient la description de la situation initiale et qui domine une complication (événements remarquables) et une résolution (réactions subséquentes des personnages), qui est suivie d’une évaluation (réactions mentaux à l’épisode) et puis d’une morale (conséquences possibles dans l’état actuel). Cette macro-structure est instanciée lors de la compréhension du texte, elle est absolument nécessaire pour la compréhension et la production des résumés. En effet, van Dijk affirme que le résumé est la réalisation concrète de la macro-structure du texte. Les macro-structures sont aussi connues pour d’autres types de textes tels que les textes d’argumentation (Sprenger-Charolles, 1992). Pour interpréter un texte et en déduire sa macro-structure, des macro-règles sont appliquées à la micro-structure. Il existe plusieurs versions des macro-règles, telle celles qu’on trouve dans (Sprenger-Charolles, 1992).

(MR1) Effacement : suppression des propositions qui ne sont pas pertinentes pour la compréhension du texte. Ainsi pour une phrase comme “Margie a un ballon rouge” qui est composée des propositions Margie a un ballon et le ballon est rouge, il est possible que la seule proposition qui reste après le processus de compréhension soit Margie a un ballon car l’attribut rouge n’est pas important dans le contexte particulier.

(MR2) Intégration : étant donné P et Q, P est intégrée dans Q si P est une condition, composante ou conséquence normale de Q. Par exemple, la séquence de propositions Une rafale captura le ballon et Margie n’a plus son ballon est intégrée dans Une rafale captura le ballon.

(MR3) Construction : P est substituée par Q si P est une composante ou une conséquence normale de Q. Par exemple, à partir de le matin Margie prend son lait, ses céréales et des œufs on peut construire Margie prend le petit déjeuner.

(MR4) Généralisation : une proposition P peut être remplacée par Q, si P est contenu dans Q (en sens ensembliste). Pour un texte comme “Margie a mangé une fraise puis une pomme et enfin une orange” il est possible que lors du processus de compréhension, la proposition Margie a mangé des fruits soit la seule qui reste de la micro-structure.

Pour un texte comme celui de Margie (Fig. 1.1) on aura un épisode qui contient une situation initiale (description de Margie et son ballon) suivie d’une complication (le vent capture le ballon), d’une résolution (le ballon crevè) et, finalement, d’une évaluation (Margie pleure). Les macro-règles expliquent parfaitement les propositions qui ont été effacées pour produire le résumé de la Fig. 1.1.

Étant donné que les règles sont assez puissantes, elles doivent être spécifiées en fonction du type de discours (narratif, d’argumentation, etc.). Toutefois, une contrainte générale pour l’application des règles est celle de la moindre généralisation : on généralise l’argument d’une proposition ou le prédicat de manière minimale. Cela explique en partie le fait que le résumé “On décrit l’état d’une fille et les causes.” pour l’histoire de Margie soit bizarre.
2.3.2 History Grammars

En particulier dans le cas des textes narratifs, Rumelhart (1975) explique la production des résumés en modélisant le processus de compréhension par des History Grammars. Elles sont définies en utilisant les éléments suivants :

(a) catégories syntaxiques : telles que Setting (description des personnages, conditions initiales, etc), Episode (événement et réaction à l’événement), Action (action d’un être humain ou d’une force naturelle), Story (discours qui est centré sur les réactions des personnages), etc.

(b) règles syntaxiques : telles que :

   Story $\Rightarrow$ Setting $+$ Episode (une histoire est composée de la description des personnages suivie d’un épisode)
   
   Episode $\Rightarrow$ Event $+$ Reaction (un épisode est composé d’un événement, puis d’une réaction à l’événement)
   
   Setting $\Rightarrow$ (States)* (la description des personnages est une séquence de zéro ou plus états)
   
   Event $\Rightarrow$ Episode $|$ Change_of_State $|$ Action $|$ Event $+$ Event (un événement est un épisode ou un changement d’état ou une action ou une séquence d’événements)
   
   etc.

(c) catégories sémantiques : telles que Allow, And, Initiate, Cause, Motivate, ...

(d) règles d’interprétation sémantique. Par exemple, la règle syntaxique Story $\Rightarrow$ Setting $+$ Episode est interprétée comme Allow(Setting, Episode) et la règle syntaxique Episode $\Rightarrow$ Event $+$ Reaction est interprétée comme Initiate(Event, Reaction).

(e) règles de condensation qui sont appliquées à la représentation sémantique, par exemple : l’information Allow(Setting, Episode) peut être condensée simplement par Episode, mettant en évidence l’importance de l’épisode central par rapport à l’introduction des personnages.

L’interprétation de l’histoire en utilisant les règles syntaxiques produit un arbre syntaxique (voir Fig. 2.9) à partir duquel on produit la représentation sémantique de l’histoire, par exemple on peut obtenir la relation sémantique Allow (“Margie serraît…”, “Tout d’un coup, une rafale…”). Ensuite, l’ensemble de règles de condensation peut être appliqué à la représentation sémantique pour obtenir le contenu essentiel de l’histoire. Ceci explique la production du résumé “Margie pleura quand le vent cassa son ballon” pour l’histoire de Margie. Les règles peuvent être appliquées à plusieurs niveaux mais il n’est pas clair comment les décisions sont prises et quelles sont les contraintes pour l’application continue des règles d’interprétation.

2.3.3 Plot Units

D’autres théories mieux motivées psychologiquement ont été élaborées dans le cadre de la science cognitive et d’intelligence artificielle. Lehnert (1981, 1984) et Lehnert et Loiselle
(1989) introduisent les *Plot Units* comme éléments pour interpréter une histoire de type narratif. Son modèle inclut l'interprétation d'un texte en *événements* (positifs et négatifs) et *états* des personnages, le rapport entre les événements et les états par des *relations causales* et, ensuite, la reconnaissance des Plot Units de haut niveau conceptuel, qui mettent en évidence des patrons thématiques (par exemple un patron comme *Competition* apparaît dans un texte comme “Avec l’argent qu’ils ont ramassé Jean veut acheter un vélo tandis que Marie veut un skate” aussi bien que dans “Paul et Pierre ont envoyé leurs C.V.s pour le poste de programmeur”). Les plot units qui sont plus centrales dans l’histoire, c’est-à-dire celles qui sont plus connectées avec le reste de l’histoire, constituent la base pour obtenir un résumé. Formellement l’histoire doit être transformée dans un graphe qui est coupé en plot units. Puis, on construit un nouveau graphe où les plot units sont des nœuds et deux plot units sont connectées si elles ont des événements en commun. Les plot units les plus connectées sont considérées comme essentielles pour le résumé. L’histoire de Margie est assez simple pour appliquer cet théorie, en effet, elle ne contient qu’une plot unit : *loss* (la perte du ballon).

### 2.3.4 Concept Coherence

Alterman (1992, 1985) et Alterman and Bookman (1990) expliquent le processus de condensation de l’information par la construction d’un graphe de concepts qui sont connectés par des relations de cohérence (*Concept/Coherence*). Les concepts sont obtenus à partir des propositions du texte ou bien par déduction. Les relations de cohérence rendent compte de rapports temporels, taxonomiques, etc. entre les concepts. Un dictionnaire contient les concepts et leurs relations qui sont restreintes en type et nombre. Si l’on considère par exemple les propositions (1) *le vent captura le ballon* et (2) *il l’entraînà vers un arbre* on peut constater que la proposition (1) est en rapport avec la proposition (3) *le vent a le ballon* par une relation de *Consequence*, mais aussi que la proposition (2) est en rapport avec (3) par une relation de *Antecedent*. Donc, on peut dire que la proposition (1) et (2) sont en rapport en utilisant la proposition (3) (qui a du être inférée pour construire la cohérence du texte). Pour l’histoire de Margie on aura une représentation comme à la
2.3. PRODUCTION DES RÉSUMÉS

Fig. 2.10. Les concepts sont les nœuds du graphe, obtenus à partir de l’histoire, et les arcs sont des relations de cohérence : Antecedent (ANT), Subclass (SC), Coordinating (COORD), Consequent (CONS), Sequence (SEQ). Les concepts qui dominent toutes les composantes de l’histoire sont pris comme base pour la construction du résumé. Ainsi, dans cet exemple les concepts Prendre, Mouvoir(1) et Frapper “résument” une partie de l’histoire de Margie.

![Diagramme des concepts et relations]  

**Fig. 2.10**: Interprétation de Margie selon les Concept/Coherence

### 2.3.5 Production chez les rédacteurs professionnels


- **facteurs d’entrée** : caractéristiques du texte source tel que type de document source et la quantité. Généralement un article technique qui traite un seul sujet favorise la production d’un résumé informatif, tandis que les actes d’une conférence ou plusieurs sujets...
sont traités favorisent des résumés indicatifs (voir exemple à la Figure 2.11). D’autres facteurs d’entrée tels que les parties du document qui méritent être résumés ont été signalés par Maizell et al. (1971);

The following topics were dealt with : recursive sets; module verification; Prolog; set theoretic reduction; existential linear theory of reals; test classes; SLDNF-resolution; information systems and domain; Davis-Putman resolution; structural complexity theory; Occam; propositional provability; polymorphic recursion; Horn classes; branching programs; predicate calculus; reducibility of monotone formulae; proof theories; data representation in lambda calculus; temporal completeness theorem; temporal logic; abstract fairness; set partitioning time complexity; SLD-resolution; logic programs; disjunctive deductive databases; and primitive recursive functions.

**Figure 2.11:** Résumé indicatif 63034 de la revue Computer & Control Abstracts pour les annales du CSL’89 3rd Workshop on Computer Science Logic

**facteurs de but** : tels que la fonction du résumé et l’utilisateur. Par exemple, si l’objectif est de signaler la parution d’un nouvel article, un résumé indicatif est suffisant tandis que si l’objectif est d’informer, alors le résumé doit être informatif. Si l’audience est spécialiste, l’information de type contextuel sera probablement omise tandis qu’elle sera pertinente si l’utilisateur est non spécialiste. Les guides de l’agence de diffusion de l’information auront une influence sur le résumé : Borko and Bernier (1975) signalent qu’une agence de diffusion de résumés du domaine de la chimie sera plutôt intéressée à résumer seulement les aspects qui portent la chimie d’un travail en biochimie. Russell (1988) indique que le type de résumé (indicatif vs. informatif) détermine le type d’information à inclure : les résumés indicatifs doivent signaler les objectifs et méthodologie de la recherche tandis que les résumés informatifs doivent aussi inclure des résultats et des conclusions. Dans le cas d’un synopsis d’un film sur un crime qui à pour but d’attirer l’attention du lecteur, il serait essentiel de ne pas dévoiler dans le synopsis l’identité de l’assassin afin d’assurer le succès du résumé. La figure 2.4 présente un synopsis du film “Psycho” où l’identité de l’assassin n’est pas dévoilée.

**facteurs de sortie** : caractéristiques du résumé par exemple des listes de mots clés, texte complet, langue dans laquelle le résumé doit être rédigé, espace disponible pour le résumé, etc.

Ces facteurs sont pris en considération afin de produire un résumé approprié pour chaque situation. Selon Cremmins (1982), pour l’étape analytique, les rédacteurs professionnels se basent sur des éléments de repère pour obtenir les informations essentielles du texte source tel que les expressions “Il est important” ou “Il est nécessaire” utilisées par les auteurs pour signaler les points relevant du document mais aussi la première section du document (normalement introduction) et la dernière section du document (normalement conclusion) sont des parties à repérer.

En ce qui concerne l’étape synthétique, les parties du texte qui ont été sélectionnées dans l’étape analytique sont transformées pour obtenir des expressions linguistiques succinctes.
(Cremmins, 1982; Mathis and Rush, 1985; Maizell et al., 1971). Tous ces aspects seront abordés plus en profondeur dans le chapitre 4 à la section 4.4.

2.4 Conclusion

Malgré la variété de types de résumé existants, la plupart des études considèrent qu'il y a deux types principaux de résumés dans le contexte du texte technique et scientifique : le résumé indicatif et le résumé informatif. Ces deux types de résumés constituent notre sujet de recherche et sont étudiés en profondeur au chapitre 4. Ici, nous avons brièvement montré comment la tâche de produire des résumés est complexe car elle nécessite la compréhension du texte source, la condensation sémantique et la reformulation. Ces processus font appel à des compétences linguistiques et sémantiques présentes chez l'être humain mais généralement difficiles à implanter de manière automatique. Dans le chapitre 4 nous montrons comment des ordinateurs n'ayant pas ces compétences peuvent produire des résumés acceptables pour une tâche donnée.
Chapitre 3

Le résumé automatique

Ce chapitre détaille quelques approches calculatoires pour la production de résumés automatiques. Étant donné la quantité d'études sur ce sujet nous nous sommes concentrés sur des approches qui montrent l'état d'avancement de la recherche et qui ont influencé le développement de notre méthode.

3.1 Introduction

L'idée de produire des résumés de manière automatique n'est pas nouvelle, la première implantation d'un système résumeur date des années 50s (Luhn, 1958). Jusqu'aux années 70s, les méthodes utilisées étaient plutôt statistiques alors que des idées issues de l'intelligence artificielle sont apparues dans les années 80s avec l'objectif de démontrer la capacité de compréhension d'un agent artificiel. De nos jours, on constate l'application d'une combinaison de plusieurs méthodes car aucune ne peut garantir un bon résultat de manière absolue.

On peut considérer deux approches générales dans la production des résumés automatiques. D'une part le paradigme d'extraction de phrases qui "semblent" contenir le contenu essentiel du document : l'objectif est de produire un extrait du document source avec lequel le lecteur puisse obtenir une idée de ce dont l'article parle. L'autre possibilité est d'utiliser des systèmes fondés sur la "compréhension" du texte source pour produire un vrai texte, soit une suite de phrases qui soit grammaticale, cohérente et cohésive. Le paradigme dominant de nos jours continue à être celui de l'extraction des phrases bien que des essais visant à produire des vrais textes existent. Quelle que soit la méthode utilisée pour produire un résumé automatique, on peut considérer que le processus de production automatique d'un résumé est composé des étapes suivantes :

Acquisition du texte : ici on inclut la transformation du texte en papier en format binaire, la segmentation du texte en différentes unités (phrases, paragraphes, etc.) et l'éventuel balisage des éléments (titres, sections, bibliographie, notes, etc.)

Interprétation du texte : qui transforme la chaîne de caractères en une représentation syntaxique, rhétorique ou conceptuelle, dépendant du cadre théorique utilisé.
 Sélection des unités : qui utilise la représentation pour décider quelles sont les unités de “contenu” les plus représentatives du texte que ce soit des mots, des phrases, des paragraphes, des sections ou des propositions.

Condensation des unités : qui élimine des propositions répétées et produit des généralisations.

Génération du résumé : qui, à partir du contenu “propositionnel” essentiel, génère un nouveau texte grammatical et respectant les caractéristiques du genre textuel.

Dans ce qui suit, le problème de l’acquisition du texte et de son balisage ne seront pas abordés. On suppose que le texte est déjà disponible en format électronique. Nous nous concentrerons sur les aspects plus centraux du processus de production du résumé automatique. De plus, on mettra en évidence les particularités de chaque étape des méthodes existantes. Pour chaque méthode nous inclurons une table qui résume les processus de interprétation, sélection, condensation et génération.

3.2 Les méthodes d’extraction de phrases

Comme on l’a déjà dit, l’objectif des méthodes d’extraction des phrases est de repérer dans le texte source les phrases les plus importantes. Le résultat obtenu est alors un extrait du texte source qui, même avec des problèmes de cohésion et cohérence, est utile pour certaines tâches.

3.2.1 Méthode de distribution de termes

Le pionnier dans le domaine de la génération de résumés est Luhn (1958) qui a introduit la méthode de distribution de termes. L’idée de cette méthode est de considérer comme “importantes” les phrases qui contiennent des mots “importants” du texte. Un mot est considéré important s’il est employé assez fréquemment dans le texte.

Le processus d’interprétation s’effectue en deux étapes. Dans la première, le texte source est traité pour calculer la fréquence de chaque mot de “contenu” du texte et dans la deuxième les fréquences sont utilisées pour associer un poids à chaque phrase.

Pour le calcul des fréquences on considère généralement les mots qui appartiennent à des classes non fermées de la langue tels que les noms et les verbes. On considère comme un même mot les mots dérivés de la même racine (par exemple, “résumé”, “résumés” et “résumer”). Le système nécessite une liste des mots qui ne doivent pas être considérés même s’ils sont assez fréquents dans un texte, par exemple prépositions, conjonctions, pronoms, déterminants, etc. Une fois la fréquence de chaque mot calculée, une liste triée par fréquence est obtenue, il s’agit de la liste de distribution de termes. Il faut noter que lorsqu’on traite un texte qui appartient à un corpus sur un sujet particulier (informatique, économie, etc.) d’autres mesures doivent être prises en considération. En effet, pour une collection de documents sur l’informatique il est très probable que des mots tels que “ordinateur” et “algorithme” soient fréquents dans tous les documents tandis que d’autres comme “système distribué” soient seulement
3.2. LES MÉTHODES D'EXTRACTION DE PHRASES

spécifiques à un sous-ensemble de documents. Dans ces cas, la fréquence de chaque mot doit être normalisée (Salton, 1988). D’abord il faut calculer la fréquence de chaque mot dans la collection complète $dt_{fi}$ (nombre de documents contenant le mot $i$), puis la fonction inverse de la fréquence, $idf_i$, est calculée (il y en a plusieurs mais en générale la formule $\log(N/dt_{fi})$ où $N$ est le nombre de documents de la collection, est utilisée) et finalement la fréquence normalisée du mot $i$ est donnée par la formule $tf_i * idf_i$ où $tf_i$ est la fréquence du mot $i$ dans le document à résumer. Seulement une partie de la liste est considérée pour l’étape suivante qui calcule le poids de chaque phrase.

Pour mesurer le poids d’une phrase, on utilise le texte source et la liste de distribution de termes. Plusieurs critères peuvent être appliqués pour calculer le poids d’une phrase : ainsi on peut considérer que le poids d’une phrase est la somme des fréquences des mots dans la phrase, la quantité de mots importants qu’elle contient, ou le fait que plusieurs mots coocurent dans la même phrase. Cette dernière approche, qui est d’ailleurs l’approche originale de Luhn, permet de considérer par exemple que la proximité des mots “base” et “données” dans l’expression “base de données” est plus importante que l’apparition isolée de chaque mot, cette heuristique met en évidence l’importance des termes. Ce processus permet d’affecter un poids à chaque phrase. Ensuite, le processus de sélection choisit les phrases les plus “pesantes”. La sélection peut être faite en terme d’un pourcentage du texte original, en nombre de phrases ou en nombre de mots.

Dans cette approche, il n’y a pas de processus de condensation des unités.

Le processus de génération consiste à juxtaposer les unités sélectionnées en ordre d’apparition dans le texte source.

La Table 4.1 résume les étapes de la méthode de distribution de termes. Les avantages de la méthode sont sa robustesse (n’importe quel texte aura un résumé) et sa facilité d’implantation. Les limitations sont toutefois nombreuses. Comme on ne prend pas en considération les relations entre les différents éléments du texte, le résultat risque d’être incohérent et même d’omettre de l’information importante.

Par exemple, dans le texte suivant :

(1) Dans cet article nous décrivons le projet X...(2) Ses objectifs sont...(3) Il vise à...

Les trois phrases parlent de l’entité “le projet X” mais étant donné que l’entité est exprimée par le biais de trois expressions linguistiques différentes (une expression définie, un pronom possessif et un pronom personnel), la méthode n’est pas capable d’identifier que les phrases (2) et (3) contiennent de l’information essentielle pour le résumé.

Toutefois dans le texte suivant :

(1) Nous présentons une méthode nouvelle pour X qui est correcte... (2) Des travaux antérieurs ont utilisé Y pour X...(3) Cette méthode était partiellement
correcte...

Bien que dans ce texte les deux occurrences du mot “métodex” aient le même sens, dans la phrase (1) on parle d’une méthode nouvelle tandis que dans (3) on parle d’une méthode antérieure. Si le système décide d’inclure les phrases (1) et (3) dans un résumé tout en oubliait la phrase (2) le résultat sera incohérent ou, pire, induira en erreur.

Un autre problème est l’ambiguïté des termes : la même forme est considérée avec le même sens, ainsi les deux utilisations de la forme “bank” en “You should ask the bank for a loan” et “I was by the river bank” sont considérées identiques.

Même avec ses limitations, cette méthode est utilisée en combinaison avec d’autres dans la plupart des systèmes actuels. Lorsqu’on considère l’approche de fréquences des mots, on risque de considérer toutes les apparitions du même mot avec le même sens et d’ignorer les relations entre les mots tels que la synonymie ; en plus, il faut y ajouter le problème des anaphores pronominales.

Pour donner plus d’importance aux concepts dont on parle dans un document, deux approches ont été récemment exploitées (à part les chaînes lexicales que l’on décrira à la section 4.2.5) : tout en exploitant une base de connaissances lexicales très riche, un processus complexe de paragraphe et de résolution des anaphores, Hahn (1990) et Hahn and Reimer (1999) arrivent à mesurer l’importance des concepts, ainsi que leurs relations et propriétés dans un domaine restreint. Ils considèrent qu’un texte fait référence à un concept aussi bien quand on le mentionne explicitement que lorsqu’on y fait référence de manière anaphorique et qu’on mentionne une propriété : par exemple, on parle du concept ordinateur quand on mentionne le concept en utilisant le mot “ordinateur”, quand on utilise l’expression “cette machine”, et quand on utilise l’expression “sa mémoire.” Des opérateurs de prédécesseurs ont été définis qui considèrent par exemple qu’un concept est important s’il a été mentionné (directement ou indirectement) plus fréquemment que d’autres. Ces opérateurs de prédécesseurs sont utilisés pour choisir parmi les concepts ceux qui ont été plus fréquemment référencés.

Lin (1995) et Hovy and Lin (1999) mesurent l’importance des concepts plutôt que des mots, et ceci afin de fusionner les concepts plus importants d’un document en les généralisant (i.e. le concept “fruit” est obtenu à partir de “pomme” et “orange”). L’importance de chaque concept est mesuré tout en considérant la mention explicite ou implicite du concept en utilisant une hiérarchie de concepts (par exemple la hiérarchie de noms de WordNet (Fellbaum, 1998)). Ainsi, le mot “fruit” compte pour une mention explicite du concept fruit et le mot “banane” compte pour une mention implicite du même concept. Étant donné que tous les noms peuvent être généralisés avec le concept “chose” des heuristiques sont appliquées afin d’obtenir des généralisations plus spécifiques. Les phrases du document source contenant les concepts retenus sont utilisées pour le résumé.

Il faut noter que de nos jours il existe des outils informatiques qui permettent de classifier avec fiabilité des mots selon leur catégorie syntaxique et que en particulier nous pouvons de manière très fiable reconnaître des groupes nominaux dans un texte quelconque. En nous
basant sur ce fait, nous utilisons la méthode de distribution pour calculer la distribution des noms comme une des heuristiques de sélection de l’information (notre mesure de pertinence est détaillée à la page 73).

<table>
<thead>
<tr>
<th>Unité textuelle</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interprétation</td>
<td>Attribuer un poids à chaque phrase en fonction de la distribution des mots dans le document source</td>
</tr>
<tr>
<td>Sélection</td>
<td>Les phrases les plus pesantes</td>
</tr>
<tr>
<td>Condensation</td>
<td>Juxtaposition</td>
</tr>
</tbody>
</table>

**Tab. 3.1: Résumé des étapes de la méthode de distribution de termes**

### 3.2.2 La méthode de la position

Cette méthode a été introduite par Edmunson (1969) pour compléter la méthode de distribution de termes qu’il a appelé “key method.” Elle est utilisée en combinaison avec d’autres méthodes d’attribution de poids pour faire augmenter ou diminuer le poids d’une phrase lors de son interprétation.

La méthode de la position considère que les premières et dernières phrases de chaque paragraphe sont importantes car elles sont considérées comme thématiques, c’est-à-dire elles “résument” le contenu du paragraphe, donc ces phrases auront leurs poids augmentés (cette affirmation est appuyée par les expériences de Baxendale (1958)). La méthode considère aussi des phrases positionnées dans certaines sections conceptuelles importantes, par exemple dans “Introduction” et “Conclusion.”

Lin and Hovy (1997) ont récemment développé des algorithmes capables d’identifier les positions thématiques dans un genre textuel quelconque (i.e., article technique, journalistique) a condition d’avoir un corpus de textes accompagnés d’une liste de mots d’ère et/ou de résumés. Les positions constituent une liste appelée *Optimal Position Policy*. Leurs expériences constituent une validation empirique de la méthode de la position.

Dans notre recherche nous avons aussi vérifié empiriquement l’importance de la position des phrases : nous montrons au chapitre 4, section 4.2.4 que certaines sections d’un article technique contiennent *a priori* de l’information pertinente pour un résumé, mais nous ne considérons la position dans la structure du texte que comme guide pour la recherche de l’information (voir page 70).

### 3.2.3 Expressions indicatives

Cette méthode a été introduit par Paice (1981) avec l’objectif de produire des résumés indicatifs. Elle est aussi connu avec le nom de “cue method.” Dans la littérature scientifique, on
<table>
<thead>
<tr>
<th>Unité textuelle</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interprétation</td>
<td>Augmenter le poids de la phrase si elle se trouve au début ou à la fin d'un paragraphe ou dans une section considéré importante (i.e., &quot;Introduction&quot;, &quot;Conclusion&quot;)</td>
</tr>
<tr>
<td>Sélection</td>
<td>Les phrases les plus pesantes</td>
</tr>
<tr>
<td>Condensation</td>
<td>Juxtaposition</td>
</tr>
<tr>
<td>Génération</td>
<td>Utilisée avec d'autres méthodes</td>
</tr>
<tr>
<td>Commentaire</td>
<td></td>
</tr>
</tbody>
</table>

Tab. 3.2: Résumé des étapes de la méthode de la position

trouve souvent des expressions qui, indépendamment du domaine particulier du texte, font référence à des catégories conceptuelles. Ainsi, si dans l'introduction d'un article on trouve une phrase avec l'expression “L'objectif de cet article est...”, on peut être presque sûr que ce qui suit dans la phrase est l'information sur les objectifs de l'article et lorsqu'on trouve une phrase qui commence par “Pour en conclure...”, on a la certitude qu'il s'agit d'une conclusion. L'idée de la méthode des expressions indicatives est de sélectionner dans tout le texte des phrases contenant ces types d'expression. Chaque expression indicative pourra avoir un poids associé et, alors, la sélection des phrases sera basée sur leur poids tel que fait Lehmann (1997).

Encore une fois, si on ne fait pas une analyse du contexte, des problèmes peuvent apparaître. Comment différencier, que dans une phrase qui commence par “La méthode a les avantages...” on fait référence à une nouvelle méthode présentée dans l'article tandis que dans une phrase comme “cette méthode a les désavantages suivants...” on parle d'une méthode qui n'a pas abouti à de bons résultats. La méthode des expressions indicatives peut être appliquée dans le domaine technique et scientifique mais il est peu probable qu'elle soit applicable dans la littérature journalistique où ce type de marqueur est assez rare.

Edmunson (1969) considère qu'il y a des expressions qui sont en soi même importantes et qui peuvent être utilisées pour identifier les phrases plus importantes d'un texte. Supposons une phrase qui commence par “Le résultat le plus significatif ...”, a priori on pourrait penser que cette phrase est “importante” indépendamment de son contenu à cause du mot “significatif”. L'idée de cette méthode est justement de considérer comme importantes des phrases contenant des mots qui, a priori, marquent des informations importantes, il appelle ces mots de “cues” et sa méthode de “cue method.” Généralement, trois listes de mots sont considérées : (i) mots pour augmenter le poids des éléments à retenir (bonus words), (ii) mots pour faire diminuer le poids d'une phrase (stigma words) et (iii) mots qui n'ont pas d'influence sur le poids de la phrase (null words). Encore une fois, l'utilisation des mots hors tout contexte peut entraîner la sélection d'information non désirée. Aussi importante dans cette approche sont les mots qui apparaissent dans les titres et sous-titres du document, cette méthode est connue sous le nom de “title method.” L'idée étant que le titre du document et les sous-titres de sections indiquent respectivement le thème et les sous-thèmes du texte. Ceci est vrai dans certains cas, mais il faut faire attention car des titres tels que “Introduction” et “Conclusion” qui apparaissent souvent dans la littérature technique sont des marqueurs
meta-textuels et des titres tels que “Méthodes et Matériaux” et “Résultats” sont des marqueurs du domaine textuel de la littérature scientifique. Dans le domaine de la médecine on trouve des sous-titres tels que “Introduction”, “Methods”, “Statistical Analysis”, “Results”, “Discussion”, “Previous Work”, “Limitations of the Study” et “Conclusion” qui marquent des sections conceptuelles et qui sont indépendantes du sujet particulier de l’article (tous les articles contiennent ces catégories conceptuelles).

La Table 4.3 résume les étapes de la méthode des expressions indicatives.

Dans notre approche qui traite du domaine technique et scientifique, nous avons fait un inventaire des expressions qui nous aident à classifier des phrases où cooccurrent des marqueurs linguistiques tels que suggérés par Paice (voir Appendices C et D). Nous utilisons à fond la structure de titres du document pour baser la sélection de l’information (voir page 90), mais nous n’attribuons pas des poids aux phrases selon les occurrences des mots des titres ou d’expressions indicatives.

<table>
<thead>
<tr>
<th>Unité textuelle</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interprétation</td>
<td>Repérer les phrases contenant des expressions indicatives préalablement définies</td>
</tr>
<tr>
<td>Sélection</td>
<td>Sélectionner un sous-ensemble des phrases repérées</td>
</tr>
<tr>
<td>Condensation</td>
<td>Juxtaposition</td>
</tr>
<tr>
<td>Génération</td>
<td>Les phrases peuvent avoir un poids</td>
</tr>
</tbody>
</table>

Tab. 3.3: Résumé des étapes de la méthode des expressions indicatives

3.2.4 Analyse rhétorique

Supposons les deux textes suivants :

(1) Chute du prix des pommes. (2) La production a augmenté dans la province cette année.

(3) Chute du prix des pommes. (4) Mangez-en !

Dans le premier texte, la proposition qui semble plus importante est la (1) parce que la proposition (2) donne les “raisons” pour la chute dans le prix (l’auteur veut nous informer que le prix est en baisse tout en donnant les raisons). Dans le deuxième texte, la proposition (3) est seulement un argument en faveur de (4) et alors (4) est plus important (l’auteur veut qu’on mange des pommes et alors il nous motive avec l’information sur la chute du prix). Entre les propositions (1) et (2), il y a une relation de cause et entre les propositions (3) et (4) il y a une relation de motivation. RST (Rhetorical Structure Theory) (Mann and Thompson, 1988) est une théorie descriptive sur l’organisation des textes qui essaie de mettre en évidence les rapports entre les propositions d’un texte. Les éléments de la théorie sont : des Relations pour décrire le rapport entre deux éléments du texte qui sont généralement
appelés **noyau** et *satellite*. Pour déduire la relation qui existe entre deux propositions il faut appliquer des jugements de plausibilité (par exemple le lecteur ne croit pas N mais si l’on dit S alors il croit N avec une certain certitude). L’autre élément qui s’ajoute à la théorie sont des *Schémas* qui spécifient la composition structurale du texte. La théorie est censée permettre la description de n’importe quel type de texte, alors elle prévoit 5 modèles de schémas qui sont montrés à la Fig. 4.1 et qui peuvent être utilisés récursivement pour décrire des textes de taille arbitraire. Les lignes horizontales indiquent des éléments du texte, les verticales ou obliques indiquent les noyaux et les lignes courbes indiquent la relation entre les éléments (marqué Rel dans la figure). Le schéma (c) qui ne contient pas de relation entre les éléments peut être utilisé pour mettre en rapport deux éléments quelconques.

![Schémas RST](image)

**Fig. 3.1: Schémas RST**

Pour analyser un texte selon la RST, il doit d’abord être segmenté en unités, en général elles sont des propositions et représentent les nœuds terminaux de l’arbre RST qui est obtenu à partir de l’application d’un certain nombre de schémas. Ainsi dans la Fig. 4.2, on présente une interprétation RST des textes suivants.

5. Jean aime les pâtes. 6. Sa famille est italienne.

7. Jean aime les pâtes. 8. Quand il va au restaurant (9) il en mange toujours un plat.

L’idée de **“nucléarité”** est centrale dans la théorie, étant donné que la plupart des relations mettent en rapport un élément qui est plus important que l’autre (le noyau), les représentations construites peuvent être utilisées pour déterminer les propositions les plus importantes du texte. Ces idées ont été implantées dans les systèmes développés par Ono et al. (1994), Miike et al. (1994) et par Marcu (1997a) visant la production des résumés.

Bien que les auteurs de la RST affirment que les relations sont indépendantes de tout marqueur grammatical (connecteur, élément lexical, etc.) les approches que l’on vient de mentionner se basent fortement sur l’existence des signaux explicites dans les textes.
Dans ces approches, l'objectif du processus d'interprétation est d'obtenir l'arbre rhétorique du texte original. Dans Ono et al., l'unité minimale d'analyse est la phrase tandis que dans Marcu elle est la clause. Des relations entre les éléments sont déduites sur la base de signaux linguistiques observés (connecteurs, etc.) et des heuristiques sont appliquées pour construire les arbres rhétoriques possibles. Le "meilleur" arbre est sélectionné sur la base de mesures de préférence.

Ce ne sont pas tous les noyaux qui sont considérés importants pour un résumé. Pour décider lesquels sélectionner il faut d'abord les mesurer. Dans l'approche de Ono et al. les phrases sont pénalisées selon leur rôle dans l'arbre rhétorique. Un poids de 1 est attribué à chaque segment satellite et un poids de 0 est attribué au noyau de chaque relation. Pour calculer le poids d'une phrase, il faut sommer les poids depuis la racine jusqu'à la phrase. Dans la Fig. 4.3 on présente un arbre RST 1 et l'assignation des poids. On a marqué dans l'arbre les noyaux (Nu) et satellites (Sat) des relations. Bien que Ono et al. considèrent que l'unité d'analyse soit la phrase, nous présentons un exemple avec des clauses pour comparer les résultats avec ceux de Marcu. Les clauses du texte sont représentées par les étiquettes $(a), (b), (c), (d), (e), (f)$.

Dans cette approche, les poids attribués à chaque clause sont : $(a) : 1, (b) : 2, (c) : 0, (d) : 1, (e) : 1, (f) : 1$. Comme le poids représente une pénalité, les clauses les moins "pesantes" seront sélectionnées pour le résumé. On obtient un ordre partiel en utilisant les poids des clauses : $(c) > (a), (d), (e), (f) > (b)$. Alors le processus de sélection doit décider l'un des ensembles de clauses suivantes :

- $\{(c)\}$ (poids 0)
- $\{(c), (a), (d), (e), (f)\}$ (poids entre 0 et 1)
- $\{(c), (a), (b), (e), (f), (d)\}$ (poids entre 0 et 2)

Ce qui donne les "résumés" :

1Il s'agit d'un exemple tiré de (Marcu, 1998)
1) They have two main advantages over magnetic stripe cards.

2) Smart cards are becoming more attractive. They have two main advantages over magnetic stripe cards. First, they can carry 10 or even 100 times as much information and hold it much more robustly. Second, they can execute complex tasks in conjunction with terminals.

3) Smart cards are becoming more attractive as the price of microcomputing and storage continues to drop. They have two main advantages over magnetic stripe cards. First, they can carry 10 or even 100 times as much information and hold it much more robustly. Second, they can execute complex tasks in conjunction with terminals.

Dans l’approche de Marcu, les unités sont promues de manière recursive. Une phrase dans le niveau n de l’arbre est promue au niveau n – 1 si elle est le noyau de la relation au niveau n – 1. Les unités placés plus haut dans l’arbre sont sélectionnées pour le résumé. Dans la Fig. 4.4, on montre comment les unités ont été promues. À coté de chaque nœud interne on a placé les unités qui ont été promues (ici les clauses ont été étiquetées (A), (B), (C), (D), (E) et (F)).
Avec cette approche on obtient l’ordre suivant : \((C) > (A) > (B), (D), (E), (F)\). Alors le processus de sélection pourra sélectionner entre les ensembles suivants :

- \(\{(C)\}\) (niveau 0)
- \(\{(C), (A)\}\) (niveau 0 et 1)
- \(\{(B), (E), (F), (A), (D), (C)\}\) (niveau 0, 1 et 2)

Donc, les résumés que l’on obtient sont :

4) They have two main advantages over magnetic stripecards.

5) Smart cards are becoming more attractive. They have two main advantages over magnetic stripecards.

6) Smart cards are becoming more attractive as the price of microcomputing and storage continues to drop. They have two main advantages over magnetic stripecards. First, they can carry 10 or 100 times as much information and hold it much more robustly. Second, they can execute complex tasks in conjunction with terminals.

Dans les deux cas, la sélection dépend d’un paramètre du système qui indique la proportion de phrases à sélectionner.

Dans cette méthode, il n’y a ni processus de condensation (pourtant étant donné que la méthode considère des clauses plutôt que des phrases, un bon niveau de réduction peut être obtenue) il n’y a pas de processus de génération à part la juxtaposition des propositions sélectionnées. La Table 4.4 résume les étapes de la méthode d’analyse rhétorique.

Les expériences avec l’utilisation de RST montrent que l’arbre rhétorique permet de prédire assez bien les unités qu’un juge humain aurait sélectionné. Mais tel qu’affirmé par Marcu, d’autres éléments doivent être ajoutés pour obtenir de bons résultats. En effet, les informations sémantiques véhiculées dans les phrases ne sont pas prises en considération pour la sélection d’éléments (elles n’apparaissent pas dans la représentation). Les textes traités par Marcu sont “assez” courts et il faudra se demander si, quand on passe du simple paragraphe au texte multi paragraphe, l’approche continue d’être applicable. Il suffit d’imaginer un article technique contenant un paragraphe avec des “travaux antérieurs” et ensuite un paragraphe qui parle de l’objectif de l’article ; la méthode calculera alors une relation joint entre les deux paragraphes et le système sera obligé d’inclure dans le résumé des informations généralement non désirées (“travaux antérieurs”). Il faut aussi se demander si l’approche est valide pour n’importe quel domaine ; Ono et al. ont montré que dans le domaine technique, les écrivains marquent les relations entre les éléments explicitement, tandis que dans d’autres domaines, les relations sont plutôt sémantiques. D’autres travaux ont utilisé la théorie RST dans le contexte du résumé automatique. Notamment, Rino and Scott (1996) utilisent une représentation sémantique et rhétorique du texte source qui est la base pour la sélection cohérente de l’information. Cependant, l’approche reste au niveau théorique car une telle représentation ne peut pas être obtenue de manière automatique.
L’analyse rhétorique appliquée dans un contexte limité comme la phrase ou le paragraphe n’a pas été explorée dans notre travail actuel mais on compte le faire dans nos travaux futurs. Une telle approche permettra d’effacer de l’information peu importante d’une phrase contenant de l’information importante. Il faut noter qu’une grammaire du discours peut aussi être utilisé pour guider la sélection de l’information pour un résumé tel que proposé par Charolles (1991).

<table>
<thead>
<tr>
<th>Unité textuelle</th>
<th>Phrase et clause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interprétation</td>
<td>Construction de l’arbre rhétorique du texte et attribution d’une pénalité à chaque unité ou promotion des unités basée sur la notion de noyau</td>
</tr>
<tr>
<td>Sélection</td>
<td>Les unités moins pesantes ou plus proches de la racine de l’arbre</td>
</tr>
<tr>
<td>Condensation</td>
<td>Juxtaposition</td>
</tr>
</tbody>
</table>

Tab. 3.4: Résumé des étapes de la méthode rhétorique

### 3.2.5 Cohésion lexicale


Salton et al. (1997) appliquent techniques de recherche d’information pour construire des résumés par extraction des paragraphes plutôt que des phrases isolées. Leur méthode se base sur l’identification de segments du texte et sur l’importance des paragraphes. Un segment est un groupe de paragraphes bien connectés, les rapports entre les paragraphes sont calculés en utilisant des mesures de similarité issues de recherche d’information (interprétation). Les paragraphes avec plus de connections sont sélectionnés pour construire le résumé (sélection). Quelques heuristiques sont appliquées pour garantir la couverture de tous les aspects du texte, c’est ici que les segments jouent un rôle majeur dans cette approche. La Table 4.7 résume les étapes de cette méthode. Les auteurs prétendent que tout en sélectionnant des
### 3.2. LES MÉTHODES D’EXTRACTION DE PHRASES

<table>
<thead>
<tr>
<th>Unité textuelle</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interprétation</td>
<td>Construire des relations de cohésion lexicale entre les phrases et classifier les phrases selon “début de thème”, “continuation de thème”, “clôture de thème” et “marginale”</td>
</tr>
<tr>
<td>Sélection</td>
<td>Sélectionner un sous-ensemble de phrases qui introduisent, continuent et terminent les thèmes</td>
</tr>
<tr>
<td>Condensation</td>
<td>Juxtaposition</td>
</tr>
<tr>
<td>Génération</td>
<td>Juxtaposition</td>
</tr>
</tbody>
</table>

**Tab. 3.5: Étapes de la méthode de cohésion lexicale**

<table>
<thead>
<tr>
<th>Unité textuelle</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interprétation</td>
<td>Construire des chaînes lexicales qui lient des phrases contenant des mots liés par des relations de cohésion lexicale</td>
</tr>
<tr>
<td>Sélection</td>
<td>Sélectionner un sous-ensemble de chaînes et ensuite un ensemble de phrases de chaque chaîne</td>
</tr>
<tr>
<td>Condensation</td>
<td>Juxtaposition</td>
</tr>
<tr>
<td>Génération</td>
<td>Juxtaposition</td>
</tr>
</tbody>
</table>

**Tab. 3.6: Étapes de la méthode des chaînes lexicales**

paragraphes les chances d’obtenir des résumés plus cohérents augmentent.

<table>
<thead>
<tr>
<th>Unité textuelle</th>
<th>Paragraphe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interprétation</td>
<td>Calculer les connections entre les paragraphes d’un texte en utilisant des mesures de similarité</td>
</tr>
<tr>
<td>Sélection</td>
<td>Sélectionner les paragraphes les plus connectés</td>
</tr>
<tr>
<td>Condensation</td>
<td>Juxtaposition</td>
</tr>
<tr>
<td>Génération</td>
<td>Juxtaposition</td>
</tr>
</tbody>
</table>

**Tab. 3.7: Étapes de la méthode d’extraction de paragraphes**

### 3.2.6 Classification des éléments

Dans les textes de science et technique il y a des phrases qui font référence à des catégories conceptuelles telles que : Connaissances Antérieures, Contenu, Méthode et Résultat, on peut également constater que dans les résumés de science et technique des informations relatives à ces catégories sont souvent retenues pour le résumé. Lehmann (1997) a développé un thesaurus d’expressions qui permettent de décider hors contexte si une phrase appartient à une
catégorie conceptuelle. Le thesaurus contient l’inventaire d’expressions qui sont associées à chaque catégorie. En plus, chaque expression à un poids qui permet de mesurer la pertinence de l’expression en tant qu’indicateur de la catégorie. L’interprétation du texte consiste en la classification de chaque phrase et dans l’attribution d’un poids, tout en utilisant le thesaurus. Le processus de sélection choisit les phrases qui ont été classifiées selon les catégories conceptuelles. Les phrases sont présentées dans l’ordre du texte source. Il n’y a pas de processus de condensation de l’information. Aussi dans le contexte du texte scientifique Teufel and Moens (1998, 1999) ont utilisé des méthodes de classification des phrases en catégories de argumentation tels que Background, Topic, Related Work, Purpose, Solution, Result, Conclusion. Pour y arriver ils utilisent des traits superficiels (i.e., information lexicale et position de la phrase dans la structure du document) et des règles de classification statistiques que nous décrirons à la section 4.3.2.

Dans l’approche de Minel et al. (2000), les phrases sont classifiées en catégories sémantiques hiérarchisées (Hypothèse, Objectif, Définition, Soulignement, etc.); pour en déduire la catégorie, un ensemble de règles d’exploration contextuelle doivent être appliquées à la phrase. Les règles attribuent une étiquette sémantique à une phrase lorsqu’un ensemble de marqueurs cooccurrent dans la phrase. Par exemple, la phrase “Il est important de souligner que...” sera classifiée avec l’étiquette Soulignement à cause de l’occurrence combinée des marqueurs “souligner”, “il-être”et “important”. Chaque phrase sera classifiée selon une catégorie (interprétation) et les phrases seront sélectionnées selon la hiérarchie des étiquettes qui indique l’importance des informations sémantiques selon l’usager du système. Une interface permet aussi d’obtenir le contexte textuel (i.e. phrase antérieure) de l’information que l’utilisateur trouve importante dans le résumé. La Table 4.8 résume les étapes de la méthode de classification sémantique.

Ces approches essayaient de classifier sémantiquement les phrases d’un texte tout en oubliant le contexte. Étant donné que l’on se base sur des marqueurs, il n’est pas possible d’assurer que la classification des éléments soit correcte. Pourtant l’approche de Minel a été évaluée par des juges humains qui ont trouvé leurs résumés automatiques de bonne qualité.

Dans notre approche, nous classifions aussi les phrases en types d’information selon la cooccurrence des marqueurs, mais nous considérons en plus le contenu de la phrase, l’importance inhérente de l’information et son rôle par rapport à l’information thématique (ceci sera élargi au chapitre 5).

<table>
<thead>
<tr>
<th>Unité textuelle</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interprétation</td>
<td>Attribuer à chaque phrase une étiquette sémantique</td>
</tr>
<tr>
<td>Sélection</td>
<td>Les phrases ayant certaines étiquettes</td>
</tr>
<tr>
<td>Condensation</td>
<td></td>
</tr>
<tr>
<td>Génération</td>
<td>Juxtaposition</td>
</tr>
<tr>
<td>Commentaires</td>
<td>Les phrases peuvent avoir un poids</td>
</tr>
</tbody>
</table>

Tab. 3.8: Résumé des étapes de la méthode de classification sémantique
3.3 Approches hybrides

Les méthodes présentées dans les sections précédentes utilisent des traits (i.e., fréquence, position, expression indicative, etc.) qu’isolément ne peuvent pas garantir des résultats optimaux. Ici nous considérons des approches qui combinent ces traits manuellement ou à l’aide d’évidences statistiques afin de produire de meilleurs résultats.

3.3.1 Combinaison manuelle

Edmunson (1969) a été le premier à combiner l’évidence de quatre méthodes : “cue” (C), “title” (T), “position” (P) et mot clés (ou “key method” (K)) en proposant l’équation \( a_1 * K + a_2 * T + a_3 * P + a_4 * C \) qui mesure le poids combiné des différentes méthodes. Les poids \( a_i \) ont été déterminé de manière expérimental. Les expériences d’Edmunson avec 200 textes ont montré que si on combine les méthodes “cue”, “title” et “position” (poids zéro pour la méthode “key”) on obtient des meilleurs résultats que si on les combine avec la “key method.” Les trois méthodes ont été reprises par Rush et al. (1971) qui les ont combinés avec une méthode d’exclusion de phrases pour produire des résumés indicatifs. L’idée est d’utiliser un dictionnaire contenant des indices sur des phrases qui ne devront jamais apparaître dans un résumé (matériel historique, exemples, spéculations, etc.).

Dans le contexte du texte journalistique, Strzalkowski et al. (1998) ont combiné la méthode de distribution de termes, la méthode du titre, la “cue method” et la position, mais ils ont aussi considéré l’évidence spécifique du genre textuel journalistique : d’une part le fait que les phrases qui contiennent de nomaux qui ont été introduit au début du texte semblent être plus pertinentes que des phrases n’ayant pas ce trait, et d’autre part le fait que des mots mentionnés dans quelques paragraphes sont plus importantes ou discriminatoires que les mots qui sont mentionnés dans tous les paragraphes (i.e. pour les identifier on utilise une espèce de mesure \( idf \) appliqué au paragraphe). Dans cette approche, la sélection se fait au niveau du paragraphe (un poids est associé à chaque paragraphe du texte) et ceci augmente les possibilités d’obtenir un résumé plus cohérent.


3.3.2 Combinaison statistique

Kupiec et al. (1995) ont introduit des méthodes de classification statistique pour le résumé automatique. L’idée est la suivante : supposons que, à partir d’un texte quelconque, on se donne la tâche de produire un extrait d’une taille donnée en nombre de phrases. A priori,
chaque phrase du texte source a la même probabilité d’être sélectionnée pour l’extrait.

Maintenant supposons que l’on a observé une “grande” quantité de textes et leur extraits (fait par des humains) et qu’on a constaté que des phrases ayant certains “traits” sont plus souvent sélectionnés pour produire l’extrait que des phrases ne les ayant pas. Par exemple, on a observé que des phrases avec des marqueurs indicatifs comme “Cet article”, “Cette étude”, “Nous concluons”, etc. sont fréquemment retenues pour l’extrait.

Maintenant, produire un extrait d’un texte qu’on n’a jamais vu peut être fait avec plus de confiance qu’avant. Si l’on trouve dans le texte une phrase avec un marqueur indicatif, alors étant donné que dans le passé des phrases avec des marqueurs indicatifs ont été souvent sélectionnés la probabilité de sélectionner cette phrase sera plus élevée. Voilà l’idée de base des modèles statistiques.

Il faut bien comprendre que l’objectif final est d’estimer la probabilité qu’une phrase avec un certain nombre de traits soit sélectionnée pour un extrait (ainsi par exemple quelle est la probabilité qu’une phrase ayant un marqueur “expression indicative” soit sélectionnée)

Donc, il faut se donner un ensemble de traits à considérer pour classifier les phrases. Par exemple on peut travailler avec les traits discrets suivants pour une phrase :

- elle a une expression indicative,
- elle figure dans l’introduction,
- elle apparaît dans la conclusion,
- elle contient un mot de contenu,
- elle est située au début d’un paragraphe,
- elle contient des mots du titre.

Ces traits sont facilement identifiables de manière automatique.

La probabilité de sélectionner une phrase $s$ pour l’extrait $E$ est donnée par la formule :

$$ P(s \in E|t_1, \ldots, t_k) = \frac{P(s \in E) \times P(t_1, \ldots, t_k|s \in E)}{P(t_1, \ldots, t_k)} \quad (3.1) $$

où :

$P(s \in E|t_1, \ldots, t_k)$ la probabilité que la phrase $s$ appartienne à l’extrait étant donné qu’on a observé les traits $t_1, \ldots, t_k$

$P(s \in E)$ la probabilité a priori que la phrase $s$ soit sélectionnée

$P(t_1, \ldots, t_k|s \in E)$ la probabilité de l’ensemble de traits $t_1, \ldots, t_k$ dans l’extrait

$P(t_1, \ldots, t_k)$ la probabilité de l’ensemble de traits $t_1, \ldots, t_k$ dans les textes
3.3. APPROCHES HYBRIDES

En général, les systèmes considèrent l’estimation suivante des paramètres :

\[
P(t_1, \ldots, t_k|s \in E) = \prod_{i=1}^{k} P(t_i|s \in E) \\
P(t_1, \ldots, t_k) = \prod_{i=1}^{k} P(t_i)
\]

Dans ces systèmes on doit ajouter à l’architecture de base un processus qui est en charge d’estimer les paramètres statistiques \( P(t_i) \) et \( P(t_i|s \in E) \) pour chaque trait à considérer dans le système. Il s’agit du processus d’entrainement. Pour le faire on a besoin d’un corpus de textes et des extraits alignés (les expériences jusqu’à présent ont utilisé des corpus d’entre 50 et 200 textes). On peut calculer autant la probabilité qu’un certain trait soit observé dans les extraits que la probabilité qu’un certain trait soit observé dans un texte quelconque en comptant chaque apparition du trait dans le corpus.

Étant donnés les paramètres du système et un nouveau texte, le processus d’interprétation doit calculer pour chaque phrase \( s \) du texte source la probabilité qu’elle soit retenue pour l’extrait étant donné qu’on a observé dans la phrase les traits \( t_1, \ldots, t_k \), pour le faire on utilise la formule (4.1). À la fin du processus, on aura que chaque phrase a une probabilité d’être sélectionnée.

Le processus de sélection choisit les phrases qui ont la plus grande probabilité.

Dans cette approche il n’y a pas de processus de condensation de l’information sélectionnée.

Le processus de génération consiste en la juxtaposition des phrases dans l’ordre du texte source. La Table 4.9 résume les étapes de la méthode probabiliste.

L’avantage des méthodes statistiques est leur robustesse, n’importe quel texte aura un extrait. On peut évaluer la performance des différents systèmes de traits et leurs combinaisons. Mais on ne peut pas garantir la cohérence du résultat. Un autre désavantage est le manque de disponibilité de corpus de textes et extraits, et même avec des corpus électroniques, il y a le problème de l’alignement des phrases du texte avec celles de l’extrait ; on doit faire normalement ces travaux à la main ce qui rend le système très couteux. Des expériences récentes montrent comment construire ce type de ressource de manière automatique (Marcu, 1999).

D’autres travaux récents ont utilisé ces méthodes de classification : Teufel and Moens (1998) étudient la classification des phrases en rôles rhétoriques ou d’argumentation (i.e., thème, résultat, objectif, conclusion, etc.) afin de produire des résumés d’articles scientifiques. Ici l’objectif n’est pas de répondre à la question : quelle est la probabilité que la phrase \( P \) soit sélectionnée pour un extrait ? mais plutôt de répondre à la question : quelle est la probabilité que la phrase \( P \) contienne de l’information sur \( C \) ? où \( C \) est une catégorie du modèle. Ils affirment qu’une telle classification peut être obtenue en considérant des traits linguistiques et de position des phrases. La classification statistique n’est pas utilisée dans notre approche, pourtant elle sera explorée dans notre travail futur, les résultats des expériences de ces chercheurs étant prometteurs.
### 3.4 Le problème de la cohésion

Lorsqu’on utilise une méthode d’extraction, on ne peut garantir que le résultat soit cohérent. Comme on l’a vu à la section 4.1, lorsque certaines phrases sont juxtaposées des problèmes peuvent apparaître tels des contradictions. Un autre problème est le suivant, supposons que le processus de sélection ait décidé de sélectionner une phrase contenant une référence anaphorique (pronom, expression définie, etc.) dont l’antécédent n’appartient pas à l’ensemble des phrases sélectionnées, le résumé devient alors illisible. Plusieurs chercheurs ont proposé des méthodes pour résoudre le problème de la sélection des phrases avec des expressions anaphoriques (Paice, 1990; Mathis and Rush, 1985). La plupart des approches proposent comme solution l’inclusion des phrases qui sont adjacentes à la phrase contenant l’anaphore de manière à réduire l’incohérence mais lorsque le résumé ainsi produit devient trop long il faut prendre une décision telle l’élimination de la phrase contenant l’anaphore. Ces approches ne cherchent pas à résoudre les anaphores mais à donner le contexte pour que le lecteur puisse le faire. Sans analyse ni syntaxique ni conceptuelle le problème s’avère assez difficile à résoudre. Dans notre approche nous ne traitons pas ce problème, pourtant comme dans la plupart des approches nous considérons certains marqueurs qui nous aident à exclure les phases avec des marqueurs anaphoriques.

### 3.5 Méthodes de compréhension et génération

À la différence des mesures “quantitatives” attribuant un poids à chaque phrase, les méthodes de compréhension et génération essaient de découvrir comment chaque phrase contribue à l’organisation du texte, quelle est la fonction de chaque phrase dans le tout. Étant donné une phrase, le système pourrait se demander si la phrase est la proposition principale ou secondaire d’un argument, s’il s’agit de la description d’un personnage dans une histoire, etc. Lorsqu’on a calculé la fonction de chaque partie du texte (phrase, paragraphe, etc.), on peut appliquer un critère pour décider quels sont les éléments à conserver pour le résumé, ainsi par exemple on pourra dire que les descriptions des personnages ne sont pas nécessaires dans les résumés. Pour produire un texte correct, il faut un module de génération qui décide comment les informations seront présentées à l’usager.
3.5. MÉTHODES DE COMPRÉHENSION ET GÉNÉRATION

Mais, pourquoi un système de production de résumés devrait-il utiliser des techniques de génération de textes? D'abord si le résumé est utilisé pour des fins d'indexation ou de classification ou si les phrases sélectionnées sont présentées avec le document original, la génération n’est pas nécessaire. Toutefois, quand le résumé est présenté à l’utilisateur indépendamment du document source (tel que dans des moteurs de recherche d’information sur le Web ou pour les services de “news digest” tel que http://cbc.ca/digest.html), la régénération devient nécessaire. Plusieurs études ont montré la nécessité de régénérer dans le domaine de la production automatique de résumés : d’abord tel que montré par Saggion and Lapalme (1998b), pour produire des résumés qui sont cohérents et cohésifs plusieurs transformations sont appliquées aux phrases qui contiennent l’information saillante du texte source. Au chapitre 4 à la page 63 nous montrons que dans un corpus de résumés produits par des rédacteurs professionnels, la plupart des phrases sélectionnées pour les résumés ont été transformées et réorganisées afin d’obtenir des expressions plus concises et pour respecter le style du genre textuel. Marcu (1999), suggère que pour construire un résumé de taille $X$ (en nombre de mots), il faut d’abord sélectionner des phrases contenant en moyenne $2.5 \times X$ mots, et ceci parce que pour produire un bon résumé il n’est pas suffisant de repérer l’information importante mais aussi de la régénérer. Jing and McKeown (1999) montrent qu’une partie significative des phrases d’un résumé humain sont produites par combinaison et réorganisation de l’information de plusieurs phrases du document source (cut-and-paste summarization).

Dans cette section nous nous concentrons sur deux approches : d’abord l’utilisation des scénarios pour le processus de compréhension, et ensuite la fouille de patrons et la génération.

3.5.1 Scénarios

Quand on lit une dépêche de presse sur un attentat terroriste, on a un certain nombre d’attentes sur l’information qui apparaîtra dans l’article. En général, on trouvera des informations sur le groupe terroriste, ses objectifs politiques (changement de la politique du pays, libération des prisonniers, etc.), les armes utilisées pour l’attentat (bombe, etc.) et les victimes.

Pour modéliser les connaissances d’une personne pour comprendre ce type de situation, plusieurs formalismes de représentation ont été proposés tels que les frames, les plans, les goals et les scripts (Shank and Abelson, 1977). Les scripts ou scénarios sont des structures de connaissance sur des suites stéréotypées d’actions dans une situation particulière. Par exemple si l’on considère une situation comme voyager en train on sait qu’il y a certains faits et événements qui sont probables tels que :

- acheter le billet (avec des pre et post conditions)
- composer le billet (avec des pre et post conditions)
- montrer le billet au chef de train
- payer une amende (si l’on n’a pas le billet ou s’il n’est pas composté)
DeJong (1982) a observé que lorsqu’on doit produire un résumé d’un événement, certains éléments de la situation seront toujours décrits tandis que d’autres ne sont pas nécessaires. Ainsi considérons le texte sur un tremblement de terre qui suit (Hutchins, 1987).

A small earthquake shook several Southern Illinois counties Monday night, the National Earthquake Information Service in Golden, Colo., reported. Spokesman Don Finley said the quake measured 3.2 on the Richter scale, “probably not enough to do any damage or cause any injuries.” The quake occurred about 7:48 p.m. CST and was centered about 30 miles east of Mount Vernon, Finley said. It was felt in Richland, Clay, Jasper, Effington and Marion Counties.

Small earthquake are common in the area, Finley said.

Un résumé possible de ce texte est :

There was an earthquake in Illinois with a 3.2 richter scale reading.

C’est-à-dire, les informations retenues pour le résumé sont : le fait qu’il y a eu un tremblement de terre, le lieu où l’événement s’est passé et la force du tremblement. Ces informations sont essentielles pour l’histoire, tandis que d’autres comme le fait que les tremblements sont habituels dans la région ne le sont pas.

Pour modéliser la capacité de résumer, DeJong a développé des sketchy-scripts, fondés sur les scripts, ce sont des structures de connaissances pour représenter les événements les plus intéressants d’une histoire. Ainsi on peut avoir des sketchy-scripts pour plusieurs situations (tremblement de terre, attentat terroriste, enlèvement, cambriolage, etc.). Pour produire le résumé d’un texte, il faut d’abord identifier le sketchy-script qui donne les attentes sur l’histoire et ensuite les utiliser pour “comprendre” tout le texte.

Dans cette approche, les processus d’interprétation et de sélection peuvent être considérés ensemble. Le texte original est examiné pour obtenir certains “cues” qui permettent d’activer un script. Ainsi par exemple un mot tel que “secouer” peut activer le sketchy-script du tremblement de terre et un mot tel que “bombe”, un attentat terroriste. Plusieurs scripts peuvent être actifs à un même temps et, à la fin du traitement, il faudra décider sur la base des slots instanciés dans les scripts quel est celui qui représente le sujet de l’histoire.

Pour le processus de génération, on utilise un modèle de résumé qui est associé au sketchy-script, les structures instanciées lors du traitement sont passées au générateur.

Il y a condensation de l’information lors de l’interprétation du texte : certaines informations d’une phrase complexe pourront être directement ignorées. Il y a aussi condensation lors de la génération quand une proposition plus conceptuelle est générée à la place des informations spécifiques sur l’événement. Ainsi pour le texte qui suit, on a condensation de l’information car la proposition du résumé n’est pas explicite dans le texte.

Marie est partie dîner. Elle s’est assise sur la table. Elle a appelé le garçon. Elle a mangé une hamburguer.
3.5. Méthodes de compréhension et génération

Un résumé acceptable est :

*Marie est allée au restaurant.*

La Table 4.10 résume les étapes de la méthode des scénarios. Pour développer un tel système il est nécessaire de disposer d’une codification des connaissances sur plusieurs domaines, ce qui n’est pas trivial. Les systèmes qui utilisent des scripts essaient toujours de faire un “match” entre le texte et l’un des sujets que le système est censé de traiter parfois avec des mauvais résultats. Bien que le système soit censé comprendre certains sujets, la technique pour y arriver est d’ignorer les parties du texte qui ne sont pas prévue par le script. Une approche censée régler cet inconvénient est celui de Tait (1982) qui essaie de mettre dans un résumé justement les informations inattendues, cette approche a été appliquée aussi dans des domaines très restreints (manger au restaurant, aller au zoo, etc.). Il nous faut ajouter que l’emphase des systèmes est surtout sur la sélection de l’information à véhiculer, les textes générés sont assez fixes en structure.

L’approche de DeJong est la base des systèmes actuels d’extraction d’information (MUC-4, 1992). Selon Gaižauskas et al. (1997), un système d’extraction d’information s’occupe de transformer le langage naturel en représentations structurées telles que les *templates* qui, instanciés, représentent un extrait de l’information clé contenus dans le texte original. Ensuite, l’information peut être enregistrée dans une base de données, utilisée pour répondre à des questions et même utilisée pour produire des résumés. Dans cette perspective, des travaux récents ont montré la possibilité d’utiliser des templates MUC instanciés par des systèmes d’extraction d’information pour régénérer un texte cohérent et cohérisé (Cancedda, 1999; Radev and McKeown, 1998). Étant donné que notre objectif n’est pas de produire un système de compréhension du langage naturel, les aspects décrits dans cette section n’ont pas eu une grande influence dans notre démarche méthodologique. Néanmoins, nous utilisons l’approche de régénération à partir de templates (voir section 5.4) que nous avons définie pour les tâches de sélection de contenu et régénération (voir pages 94 à 101).

<table>
<thead>
<tr>
<th>Unité textuelle</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interprétation et</td>
<td>Identification d’un script et traitement en fonction des</td>
</tr>
<tr>
<td>Sélection</td>
<td>attentes</td>
</tr>
<tr>
<td>Condensation</td>
<td>Propositions déduites à partir des propositions du texte</td>
</tr>
<tr>
<td>Génération</td>
<td>Modèle associé au script identifié</td>
</tr>
</tbody>
</table>

Tab. 3.10: Résumé des étapes de la méthode des scénarios

3.5.2 Instanciation de patrons et génération

Supposons que l’on doive faire des résumés pour des articles d’un domaine textuel restreint (Paice et Jones, 1993) et qu’on sache que :

- les articles dont on doit faire les résumés parlent toujours de concepts et des relations bien définies dans le domaine ;
- les formes linguistiques utilisées pour exprimer autant les concepts que les relations sont limitées ;
- les résumés à produire doivent toujours décrire un certain nombre de concepts et de relations.

Ces hypothèses sont assez vraisemblables, en effet, une étude menée par Liddy (1991) sur les résumés du domaine de la recherche empirique a montré que certains concepts abstraits à inclure dans un résumé et leurs relations doivent toujours apparaître dans le résumé lorsqu’ils se trouvent dans le texte source.

Si l’on prend, par exemple, le domaine de l’agriculture on peut constater qu’on parle des concepts suivants :
- SPECIES
- PROPERTIES
- PEST
aussi qu’il y a des relations entre ces concepts telles que :
- \( \text{est._une._parasite._de}(PEST, SPECIES) \)
- \( \text{est._une._propriete._de}(PROPERTY, SPECIES) \)

Lorsqu’on trouve une phrase comme “A.lolii is a common pest of ryegrass” on peut déduire que “A.lolii” est une PEST, que “ryegrass” est une SPECIES et aussi que la relation entre ces deux concepts est \( \text{est._une._parasite._de}(“A.lolii”, “ryegrass”) \). On peut repérer cette phrase du texte avec le modèle “PEST is a? pest of SPECIES” où :
- PEST et SPECIES sont les variables du modèle à être instanciées
- ? indique un segment quelconque de texte entre les mots “a” et “pest”

Si l’on est capable de faire un inventaire des modèles qui sont utilisés pour exprimer les concepts et les relations dans le domaine, on pourra les utiliser pour chercher l’information dans le texte source et ensuite l’exprimer dans un résumé. L’inventaire de modèles doit être fait à partir de l’analyse d’un corpus.

On peut identifier un seul processus qui est en charge de l’interprétation et de la sélection de l’information. Le texte est traité pour chercher tous les modèles possibles. Lorsqu’un modèle est trouvé, ses “slots” sont instanciées avec le matériel textuel contre lequel on a fait l’appariement. En cas d’ambiguïté, des heuristiques sont appliquées pour décider l’instanciation d’un slot.

Pour le processus de génération, on utilise un ensemble de patrons avec lesquels on décrit les relations entre les concepts. Ainsi on pourrait avoir un patron comme “This paper studies the effect of X on Y” où X et Y sont des variables à être instanciées. Avec lui on peut produire les phrases suivantes : “This paper studies the effect of the pest G.paluda has on the yield of potato” et “This paper studies the effect of soil temperature on the emergence of winter wheat.” Cette approche permet d’obtenir des résumés indicatifs, car la sélection de
matériel de type résultat et conclusion est beaucoup plus difficile. La Table 4.11 résume les étapes de la méthode de patrons.

Ceci est l'une des seules approches qui essaient de générer un nouveau texte, mais elle reste très limitée tant dans le type de résumé généré que dans le domaine traité. La limitation du domaine est nécessaire afin de pouvoir définir les concepts et les relations entre les concepts. L'implantation du modèle est assez simple une fois qu'on a une description formelle du domaine et l'inventaire de modèles.

Tout comme dans cette approche, nous avons exploré l'utilisation de modèles et l'instanciation de patrons. Pourtant, notre méthode a été développée indépendamment d'un domaine particulier. Tous ces aspects sont détaillés aux chapitres 5 et 6.

<table>
<thead>
<tr>
<th>Unité textuelle</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interprétation et Sélection</td>
<td>Fouille des patrons préalables dans les textes</td>
</tr>
<tr>
<td>Condensation</td>
<td>Informations répétées ignorées et extraction des fragments de phrases</td>
</tr>
<tr>
<td>Génération</td>
<td>Plusieurs modèles fixe de résumé associé au domaine textuel</td>
</tr>
</tbody>
</table>

Tab. 3.11: Résumé des étapes de la méthode des patrons

3.6 Conclusion

Dans ce chapitre nous avons exploré quelques approches qui ont été proposées pour résoudre le problème de produire des résumés de manière automatique. Il semble bien que les approches purement statistiques, simples à implanter et rapidement adaptables à d'autres domaines sont arrivées à la limite de leur puissance. D'autre part les approches symboliques fondées sur des connaissances syntaxiques, lexicales et du monde, ne peuvent être appliquées que dans des domaines restreints où l'acquisition des connaissances à tous les niveaux se fait presque toujours à la main, sauf pour l'approche de Rau et al. (1989), rendant l'adaptation à un nouveau un domaine très coûteuse. L'avenue à explorer dans le domaine de la génération de résumés est celle de la combinaison de différentes méthodes et l'incorporation d'outils robustes de traitement du langage naturel et d'acquisition automatique des connaissances (apprentissage par machine) car aucune ne garantit des résultats optimaux. Quelques groupes ont commencé à paver cette avenue (Lin, 1998; Hovy and Lin, 1999) avec la combinaison des connaissances symboliques (analyseurs syntaxique rhétorique et bases de données lexicales) et statistiques (distribution de termes et positions thématiques), mais il reste encore beaucoup de travail à faire pour en tirer des conclusions. Des nouvelles tendances dans la génération des résumés sont la considération de l'utilisateur du résumé (Tombros et al., 1998) et aussi la production des résumés de plusieurs articles qui traitent du même sujet.
(Radev and McKeown, 1998; Barzilay et al., 1999).

Afin de produire un modèle bien fondé, nous nous basons sur l’étude d’un corpus. Néanmoins, pour l’implanter nous utilisons plusieurs techniques qui ont été présentées dans ce chapitre.
Chapitre 4

Observations from the Corpus

In this chapter, we explore human produced abstracts in order to answer the following questions: where does the information reported in abstracts come from? how can it eventually be found in source documents? what is its status? and how is it conveyed in the abstract? The answer to these questions will provide a well-founded basis for the definition of a method for automatic text summarization of technical texts.

4.1 Human Produced Abstracts

Abstracts of technical articles are produced by their author or by professional abstractors working for abstracting services. These professionals do not need to understand the whole document in order to identify what the document is about. They use specific strategies in order to grasp the essential content of the document to further write their abstracts by reusing already stated conceptual information, and eventually using standard patterns of text production (Endres-Niggemeyer et al., 1995). Most studies in Information Science agree on a two stage logical account for describing the human production of abstracts: the analytical stage in which the salient facts of the text are obtained and condensed and the synthetic stage in which the text of the abstract is produced (Pinto Molina, 1995). The question is how to identify the salient information in a technical article and how to express that information in a concise text.

While all the concepts of a technical article contribute somehow to the message of the text, not all of them constitute the topic, subject or theme of the document. The topics are those concepts the author talks more about, the question is which specific information about them has to be brought in the abstract. As abstracts report the essential content of their source documents, the study of the relations between the source and the abstract can bring out conceptual and linguistic evidence for the identification of the topics.

Abstracting manuals (Borko and Bernier, 1975; Cremmins, 1982; Rowley, 1982) specify that the information to be reported in an abstract of a technical article refers to the “purpose”, “method”, “results”, “conclusions” and “findings” of the work, but unfortunately, little indication about the procedures to use to produce such classification of the information is
given. In the case of computer science abstracts, Maeda (1981) identifies four main types of information generally reported: “theme”, “method”, “results” and “discussion” of the research work; again, no information is available about the particular patterns used to interpret the text in the given framework. Liddy (1991) studied types of information in abstracts of empirical research and showed that the material reported in the abstract has a well-defined informational status comprising 36 types of information grouped in a three-level hierarchy. The most important information in this text-type refers to “purpose”, “hypothesis”, “methodology”, “subject”, “results”, “conclusions” and “references” and these categories constitute the first level of her model called Prototypical Structure. The other two levels are the Typical Structure and the Elaborated Structure which include less frequent information. She also indicates lexical clues that can eventually be used in order to classify the information. To our knowledge, this is one of the most complete work in this field and it inspired our research.

4.2 Where is the Topic?

4.2.1 Corpus Description

Our corpus is composed of 100 items, each composed of an abstract produced by a professional and its source (or parent) document. We used as source for the abstracts the journals Library & Information Science Abstracts (LISA), Information Science Abstracts (ISA) and Computer Abstracts. The parent documents were found in journals of Computer Science (CS) and Information Science (IS) such as AI Communications; AI Magazine; American Libraries; Annals of Library Science & Documentation; Artificial Intelligence; Computers in Libraries; IEEE Expert; among others (a total of 44 publications where examined, see Appendix A). The professional abstracts contain 3 sentences on the average, with a maximum of 7 and a minimum of 1. The source documents cover a variety of subjects from IS and CS. We examined 62 documents in CS and 38 in IS, some of them containing author provided abstracts. Regarding the form of the documents, most of them (97) are structured in sections; but apart from conceptual sections such as “Introduction” and “Conclusion” they do not follow any particular style (articles from medicine, for example, usually have a fixed structure like “Introduction”, “Method”, “Statistical Analysis”, “Result” “Discussion”, “Previous Work”, “Limitations”, “Conclusion” but this is not the case in our corpus). The documents are 7 pages on the average, with a minimum of 2 and a maximum of 45. Neither the abstracts nor the source documents were electronically available, so the information was collected through photocopies, thus, we do not have information regarding number of sentences and words in the source document.

4.2.2 Corpus Analysis

We manually aligned each sentence of the professional abstract with one or more elements of the source document. This was done by looking for a match between the information in the professional abstract and the information in the parent document. The structural parts of the parent document we examined are: the title of the parent document, the author
abstract, the first section, the last section, the subtitles and captions of tables and figures. When the information was not found, we looked in other parts of the parent document. The information is not always found in the source document, in which case we acknowledge that fact. This methodological process was established after studying procedures for abstract writing (Crenmins, 1982; Rowley, 1982) and some initial observation in our corpus.

For each element of the corpus, we construct a table of alignment that associates the information of the abstract with the matched information in the source document. The tables contain three columns: the first column contains the sentences of the professional abstract, the second one contains the matched information from the source document and the third column contains a label which specifies in which particular location of the source document the information was found (1st/Introduction, Last/Conclusion, -/Abstract, -/Title, -/Subtitle, -/Captioning, etc.).

In the following we will present three complete examples from the corpus, and then several examples of alignments from different items of the corpus. These are good representatives of the different situations observed and give a general idea about the distribution and use of the information.
### 4.2.3 Tables of Alignment

In Table 4.1, we show an item from the corpus. The 3 sentences of the professional abstract were aligned with 4 elements of the source document, 2 in the introduction and 2 in the author provided abstract. In this example the information of the abstract was “literally” found in the source document.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
<th>P/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presents the results of an empirical study that investigates the movement characteristics of a multi-modal mouse - a mouse that includes tactile and relevance feedback.</td>
<td>In this paper, we present the results of an empirical study that investigates the movement characteristics of a multi-modal mouse - a mouse that includes tactile and force feedback.</td>
<td>1st/Intr.</td>
</tr>
<tr>
<td>Uses a simple target selection task while varying the target distance, target size, and the sensory modality.</td>
<td>Our experiment used a simple target selection task while varying the target distance, target size, and the sensory modality.</td>
<td>1st/Intr.</td>
</tr>
<tr>
<td>Significant reduction in the overall movement times and in the time taken to stop the cursor after entering the target were discovered, indicating that modifying a mouse to include tactile feedback, and to a lesser extent, force feedback, offers performance advantages in target selecting tasks.</td>
<td>We found significant reductions in the overall movement time and in the time to stop the cursor after entering the target.</td>
<td>-/Abs.</td>
</tr>
<tr>
<td></td>
<td>The results indicate that modifying a mouse to include tactile feedback, and to a less extent, force feedback, offers performance advantages in target selection tasks.</td>
<td>-/Abs.</td>
</tr>
</tbody>
</table>

Tab. 4.1: Alignment of the professional abstract LISA 3024 and the source document “Movement characteristics using a mouse with tactile and force feedback” International Journal of Human-Computer Studies, 45(5), Oct'96 p483-93

The information of the first sentence of the abstract, which introduces the theme of the document, is found in the introduction of the source document. We observe differences in the expression of the verbs, Presents vs. We present, and also the presence of the marker In this paper, in the source document.

In the second sentence, again stylistic differences are noted: personal style in the source document, Our experiment used, and impersonal style in the abstract, Uses.

The information of the third sentence was found in two sentences of the author provided abstract. Here, the abstractor changed the verbs, find into discover and the personal style into impersonal, We found into were discovered, and also merged the information in one sentence.

This alignment shows that the organization of the information in the abstract does not always mirror the organization of the source document.
4.2. WHERE IS THE TOPIC?

In Table 4.2, the three sentences of the professional abstract were aligned with three sentences from the introduction of the source document. In this case, the information was “literally” found in the source.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
<th>P/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>The next generation of fuzzy expert systems will combine techniques such as logic programming, fuzzy theory, and neural networks to improve performance and productivity.</td>
<td>The next generation of fuzzy expert systems will combine techniques such as logic programming, fuzzy theory, and neural networks to improve performance and productivity.</td>
<td>1st/Intr.</td>
</tr>
<tr>
<td>Introduces hybrid architecture and describes combination and fusion types.</td>
<td>After a brief introduction to hybrid architecture, we describe two types: combination and fusion.</td>
<td>1st/Intr.</td>
</tr>
<tr>
<td>Discusses the development and implementation of a combination architecture fuzzy system for steel making plant, and considers the significant algorithmic strength that fusion architecture lends fuzzy systems.</td>
<td>We discuss the development and implementation of a combination architecture fuzzy system for steel making plant, and consider the significant algorithmic strength that fusion architecture lends fuzzy systems.</td>
<td>1st/Intr.</td>
</tr>
</tbody>
</table>

Tab. 4.2: Alignment of the professional abstract LISA 4600 with the source document “Fuzzy and neural hybrid expert systems: synergic AI”, M. Funabashi and others. IEEE Expert, 10(4) Aug 95, p32-42

In the first aligned sentence no differences are noted. Some differences are noted in the second sentence of the abstract: the impersonal style in the abstract, describes, vs. the personal style in the source document, we describe; the use of the verb Introduces in the abstract instead of the noun introduction, and the elision of the marker After (breaking the sequentiality of the information given by that marker); and the restatement of the nominal (combination and fusion types from the original formulation two types: combination and fusion).

Finally, in the third sentence of the abstract, we observe impersonal style in the abstract, Discusses, and personal style in the source, We discuss.
The information of the professional abstract on Table 4.3 was found in different sections and structural parts of the source document (first, second, forth, sixth and last sections and subtitles).

The information of the first sentence of the professional abstract was found in the first and last sections of the source document. Note that the abstract contains less information than the matched sentences: the objectives of the CEC Libraries Programme, which are explicit in the source, are not reported in the abstract.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
<th>P/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>The DECIMAL project is 1 of 4 projects supported under the commission of the European Communities' Libraries Programme's CAMILE (Concerted Action on Management Information for Libraries in Europe) project.</td>
<td>DECIMAL is one of four RTD Project supported under the Commission of the European Communities (CEC) Libraries Programme (Framework 3, Action Line IV, Theme 19 bis) for the development of models and tools to support decision-making in libraries.</td>
<td>1st/Intr.</td>
</tr>
<tr>
<td>It aims to produce an integrated decision support module for library management systems.</td>
<td>CAMILE is a CEC Concerted Action which commenced during autumn 1996, involving all the Partners in the four existing Projects.</td>
<td>Last/-</td>
</tr>
<tr>
<td>It is being developed from an assessment of the needs and practices of library managers in small to medium size libraries in the UK, Spain and Italy.</td>
<td>The aim of the DECIMAL (DECision-Making in Libraries) Project is to produce a commercially viable integrated Decision Support Module for library management systems, developed from a fundamental assessment of the needs and practices of library managers in small to medium size libraries in the UK, Spain and Italy.</td>
<td>2nd/-</td>
</tr>
<tr>
<td>It began in Feb 95, and comprises 4 phases: management, research, technical development and evaluation.</td>
<td>The DECIMAL project commenced in February 1995 and will run for two years.</td>
<td>4th/-</td>
</tr>
<tr>
<td></td>
<td>The project involves three practical phases (in addition to a management 'phase') : research, technical development and evaluation.</td>
<td>4th/-</td>
</tr>
</tbody>
</table>

Tab. 4.3: Alignment of the professional abstract LISA 6863 with the source document “The DECIMAL project: decision-making and decision support in small to medium size libraries” T. Oulton and others. Vine, (103) 1996, p13-19. (continues on Table 4.4)
4.2. WHERE IS THE TOPIC?

The information brought in the second and third sentence of the abstract was found in one sentence of the source, in the second section of the document. The source contains more information than the abstract: the expression *a commercially viable integrated decision support module* is more detailed than the expression *an integrated decision support module*. The linguistic expression used to introduce the objectives and development of the *DECIMAL project* is also different in source and abstract (It aims to instead of The aim of X is, and It is being developed instead of developed).

The information in the fourth sentence was found in two sentences from the fourth section. The duration of the project is omitted and some stylistic variations are also noted (use of the pronoun *it* in the abstract and change of the verb *commence* to *begin*).

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
<th>P/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Describes its objectives and structure and details progress in the technical development to date and a summary of the funding of the research phase.</td>
<td>The article describes the aims, objectives and structure of the DECIMAL project and gives an account of progress to date with a summary of the findings of the Research phase.</td>
<td>1st/Intr.</td>
</tr>
<tr>
<td>The module is being developed to incorporate both textual and numeric information to support the decision process.</td>
<td>Development of the decision support module</td>
<td>-/Subt.</td>
</tr>
<tr>
<td></td>
<td>A key focus of the technical specification was that of data types and data structures to accommodate and integrate operational data generated from library management systems with data internal or imported from external sources.</td>
<td>6th/-</td>
</tr>
<tr>
<td></td>
<td>Such data may be numeric or textual.</td>
<td>6th/-</td>
</tr>
</tbody>
</table>

Tab. 4.4: (Continues from Table 4.3) Alignment of the professional abstract LISA 6863 with the source document “The DECIMAL project: decision-making and decision support in small to medium size libraries” T. Oulton and others. Vine, (103) 1996, p13-19

Regarding the fifth sentence, its content was found “literally” in the introduction with the following differences: the expression *Describes* in the abstract vs. *This article describes* in the source document, the expression *its objectives and structure* in the abstract vs. *objectives and structure of the DECIMAL project* in the source document, and the use of the verb *detail* in the abstract instead of the original *give*.

Finally, the information of the last sentence is found, but not literally, in subtitles and sixth section of the source. Again in this example, the structure of the abstract does not mirror that of the source.
Other examples of aligned sentences which come from several items of the corpus are shown in Table 4.5. They give an insight about the alignments of sentences in the abstract with each type of structural element in the source document which will be used for reference in the following sections.

<table>
<thead>
<tr>
<th>Ex.</th>
<th>Professional Abstract</th>
<th>Source Document</th>
<th>P/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Presents a more efficient Distributed Breadth-First Search algorithm for an asynchronous communication network.</td>
<td>Efficient distributed breadth-first search algorithm</td>
<td>-/Title</td>
</tr>
<tr>
<td>(2)</td>
<td>Describes the work of the Danish Library Centre (DBC), a service organization for the entire Danish library system.</td>
<td>The Danish Library Centre. A service organization for the entire Danish library system</td>
<td>-/Title</td>
</tr>
<tr>
<td>(3)</td>
<td>Summarizes and make suggestions for future research.</td>
<td>Summary and future research</td>
<td>-/Subt.</td>
</tr>
<tr>
<td>(4)</td>
<td>Gives a formal description of the problem of optimal pruning of decision trees.</td>
<td>Pruning a decision tree is the process of replacing some of the subtrees of DT by leaves.</td>
<td>1st/Intr.</td>
</tr>
<tr>
<td>(6)</td>
<td>Shows how relation algebras can be used to handle interval reasoning.</td>
<td>The idea in this paper is to see how relation algebras can be used to handle interval reasoning.</td>
<td>1st/Intr.</td>
</tr>
<tr>
<td>(7)</td>
<td>Analyzes the complexity of the algorithm, and gives some examples of performance on typical networks.</td>
<td>We analyse the complexity of our algorithm, and give some examples of performance on typical networks.</td>
<td>1st/Intr.</td>
</tr>
<tr>
<td>(8)</td>
<td>Investigates the phase transition in binary constraint satisfaction problems.</td>
<td>The phase transition in binary constraint satisfaction problems, i.e. the transition from a region in which almost all problem have many solutions to a region in which almost all problems have no solutions, as the constraints become tighter, is investigated.</td>
<td>-/Abs.</td>
</tr>
</tbody>
</table>

Tab. 4.5: Alignments of Differents Sentences from the Corpus (continues on Table 4.6)
4.2. WHERE IS THE TOPIC?

<table>
<thead>
<tr>
<th>Ex.</th>
<th>Professional Abstract</th>
<th>Source Document</th>
<th>P/T</th>
</tr>
</thead>
</table>
| (9) | A tesseral temporal reasoning system has been designed, based on tesseral addressing and using tesseral arithmetic.  
It offers the advantage that it is compatible with existing GIS technology. | This has resulted in a tesseral temporal reasoning system, based on tesseral addressing and using tesseral arithmetic, which offers the advantage that it is directly compatible with existing GIS technology. | 1st/Intr. |

Tab. 4.6: (Continues from Table 4.5) Alignments of Different Sentences from the Corpus

4.2.4 Distributional Results

The 309 sentences of the professional abstracts were manually aligned with 568 elements in the source document. We were not able to align 6 sentences of the professional abstracts. Other studies have already investigated the alignment between sentences in the abstract and sentences in the source document. Kupiec et al. (1995) report on the semi-automatic alignment of 79% of sentences of professional abstracts in a corpus of 188 documents with professional abstracts. Using automatic means it is difficult to deal with conceptual alignments that appeared in our corpus such as the relation we show in example (4) of Table 4.5. Teufel and Moens (1998) report on a similar work but this time on the alignment of sentences from author provided abstracts. They use a corpus of 201 articles obtaining only 31% of alignable sentences by automatic means. No information is given about the distribution of the sentences in structural parts in the source document.

<table>
<thead>
<tr>
<th>Documents</th>
<th>with A.Abs.</th>
<th>w/o A.Abs.</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
<td>#</td>
</tr>
<tr>
<td>Title</td>
<td>10</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Author Abstract</td>
<td>83</td>
<td>15</td>
<td>83</td>
</tr>
<tr>
<td>First section</td>
<td>195</td>
<td>34</td>
<td>61</td>
</tr>
<tr>
<td>Last section</td>
<td>18</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Subtitles &amp; Capt.</td>
<td>191</td>
<td>33</td>
<td>76</td>
</tr>
<tr>
<td>Other sections</td>
<td>71</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>568</td>
<td>100</td>
<td>245</td>
</tr>
</tbody>
</table>

Tab. 4.7: Distribution of Information

In Table 4.7, we present the distribution of the sentences in the source documents which were aligned with the professional abstracts in our corpus. We consider all the structured documents of our corpus (97 documents). The first three columns contain respectively the information for all the documents, for documents with author abstract and for documents without author abstract (the information is given in total of elements and in percent). The
last column indicates the average of the information. We found that 72% of the information for the analytical stage comes from the following structural parts of the source document: the title of the document, the first section, the last section and the subtitles and captions. It is important to note that even in the case of documents with author provided abstract, information from other parts of the source document would be used in the professional abstract as Table 4.7 shows. In addition, a “typical” element of the corpus (average information) will contain 40% from the introduction. Our findings are in agreement with the study of Endres-Niggemeyer et al. (1995). They found that in order to produce the “topical” sentence of an abstract the professional abstractor will use the introduction and conclusion of the source document.

4.2.5 Analysis of the Results

The results so far indicate a correlation between the information in the abstract and structural parts of the source document. But it is necessary to understand why some information is considered important for the abstract, how to identify the information in the source document and how to use it in order to produce an abstract.

Sharp (1989) reports on experiments carried out with abstractors were it is shown that introductions and conclusions provide a basis for producing a coherent and informative abstract. In fact abstractors use a “short-cut” strategy (looking at introduction and conclusion) prior to regarding the whole paper. But our results indicate that using just those parts is not enough to produce a good informative abstract. Important information is found in sections other than the introduction and conclusion. Our observations regarding the use of structural parts of the source document are as follows (refer to Tables 4.5 and 4.6 for the alignments):

- Titles of articles usually contain complete descriptions of the themes of the document so they could be used by the abstractor in order to convey the information in a more precise form. Examples of such a situation are alignments (1) and (2) where the complete descriptions of the title is used in a sentence.

- Usually, statements related to the purpose, the method, and the problem studied in a technical article are found in introductions. These conceptual categories are generally reported in abstracts and can thus be extracted from the introduction because they are lexically marked. In sentence alignment (1) the objective or topic of the paper is stated in the introduction. In the alignment (7) the author states the plan of the document.

- In the conclusion, the author usually restates the objective or main topic of the document, this case is exemplified in sentence alignment (1).

- Subtitles of sections usually indicate the sub-themes of the document and also complete descriptions of entities. In sentence alignment (3) the subtitle indicates the content of the section so it is used in the abstract to indicate that information.

- As tables (and figures) usually convey information about the results of an investigation, their captions could be used to indicate the content reported in tables as sentence alignment (5) shows.
4.2. WHERE IS THE TOPIC?

- An important part of the abstracts came from other sections of the document: one example is given in the alignment presented in Tables 4.3 and 4.4, where the details about the project being described in the article are stated. In that case the information is relevant because it refers to the main entity being described in the document and not because of any lexical marker.

- Sometimes author provided abstracts are less structured than professional abstracts but abstractors use the information found in the author abstract because it is clearly stated. As sentence alignment (8) shows, the information about the investigation is stated in the author provided abstract and used in the professional abstract.

Abstractors not only select the information for the abstract because of its particular position in the source document, they also look for specific types of information which happen to be lexically marked. Having no access to the abstractors' working environment, we can just make the following observations about the characteristics of the extracted information.

We found that the information extracted by professionals contains lexical markers of relevance of a topical entity (“The principal distinguishing features of EQLIPSE are...”), theme or topic of the paper (“The subject of this paper is the concept of descriptor equivalence and...”), purpose of the article (“The purpose of this paper is to assess retrospectively...”), main conclusion (“Our conclusion was that simple and local transformations can be automated...”), results of the work (“We found that significant...”) and plan of the document (“we will first put the Word Manager project in perspective...we will then describe the progress made...”) among others (see Appendix B for the initial list of markers).

In fact 205 aligned sentences from the main sections of the source documents in our corpus contain strong lexical markers referring to concepts and relations which are typical of technical papers. This represents 35% of the total of the aligned sentences. If, in addition, we take into account the fact that 35% of the aligned elements come from titles and subtitles, we obtain that 70% of the information is somehow “indicative” of the content of the documents. In summary, looking for text spans containing “indicative expressions” and using titles and subtitles when they clearly mark the themes being described is a good strategy for grasping the content of a text.

The results of our analysis indicate that it is useful to look for information in specific parts of the source document using lexical markers to obtain part of the information for the abstract but, the material obtained in this manner is sometimes too indicative. In order to produce a good informative abstract, the information from introductions and conclusions have to be expanded somehow. Which information to expand depends in part on the reader's interests: if the abstract is to be used as a decision tool, an indicative abstract is useful but if the reader wants more information about the entities being described in the document, then some additional information has to be obtained from other sections of the document and integrated as a coherent whole.
4.3 Conceptual and Linguistic Model

The scientific and technical article is the result of the complex process of scientific inquiry (Bunge, 1967) that starts with the identification of a problem and ends with its solution. It is a complex linguistic record of knowledge referring to a variety of real and hypothetical concepts and relations. Some of them are domain dependent (diseases and treatments in Medical Science, atoms and fusion in Physics; algorithms and proofs in Computer Science) while others are common to all the technical literature (authors, the research article, the problem, the solution, etc.).

In Table 4.8, two alignments are presented.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reports work comparing speech and keying in the context of a (simulated) command and control application.</td>
<td>In this paper, we report work comparing speech and keying in the context of a simulated command and control application.</td>
</tr>
<tr>
<td>The performance of a particular two-connected mesh network, the Manhattan street network (MS.Net), is compared with that of the distributed queue dual bus (DQDB) network, which has been standardized by the IEEE 802.6 committee for MANs.</td>
<td>In this paper, we compare the performance of a particular two-connected mesh network, the Manhattan street network [4,5] (MS.Net), with the DQDB network, that has been standardized by the IEEE 802.6 committee for MANs [6].</td>
</tr>
</tbody>
</table>

**Tab. 4.8: Concepts in Source Documents and Professional Abstracts**

Ignoring for the moment differences in linguistic expression, it is clear that in the first sentence alignment of Table 4.8, the sentence of the source document is introducing one specific type of information: the theme or topic of the document and that the evidence comes from lexical items such as *this paper*, a strong reporting verb like *report*, and the explicit mention of the authors, *we*. In the second sentence alignment, the main theme is also introduced but this is done by specifying what the author studies in the document, another different type of information. The evidence for that comes from the lexical item *compare*, a strong verb associated with that cognitive activity, and the explicit mention of the authors, *we*. Note that in the first example, the verb *compare* is also used but not as the main verb of the sentence. Those two types of information are independent of any domain because they refer to normal activities done by authors of research papers: looking for a problem, studying, investigating, thinking, concluding, reporting, describing, etc.

We have identified 55 concepts and 39 relations, which are typical of a technical article, relevant for the task of identifying types of information for text summarization. This was done by the process of collecting domain independent lexical items and linguistic constructions from the corpus and classifying them using a thesaurus (Vianna, 1980). Afterwards, we expanded the initial set with more valid linguistic constructions not observed in the corpus.
4.3. CONCEPTUAL AND LINGUISTIC MODEL

Concepts can be classified in categories such as: the authors (the authors of the article, their affiliation, researchers, etc.), the work of the authors (work, study, etc.), the research activity (actual situation, need for research, problem, solution, method, etc.), the research article (the paper, the paper components, etc.), the objectives (objective, focus, etc.), the cognitive activities (presentation, introduction, argument, etc.). Some of these concepts are presented in Table 4.9 while the complete specification is presented in Appendix C.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Explanation &amp; Example</th>
<th>Lexical Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>author</td>
<td>The authors of the article. “I refer to ...”</td>
<td>we, I, author, us</td>
</tr>
<tr>
<td>author related</td>
<td>Authors’ related entity. “The core of our system is comprised of...”</td>
<td>our, my</td>
</tr>
<tr>
<td>research paper</td>
<td>The technical article “In this article...”</td>
<td>article, here, paper</td>
</tr>
<tr>
<td>research</td>
<td>The research work. “... a broad range of scientific research...”</td>
<td>research, investigation</td>
</tr>
<tr>
<td>problem</td>
<td>The problem under consideration “The lack of a library severely limits the impact of...”</td>
<td>difficulty, issue, ...</td>
</tr>
<tr>
<td>need</td>
<td>A necessity. “...the need for an interface between ...”</td>
<td>need, necessity, ...</td>
</tr>
<tr>
<td>acronym</td>
<td>An acronym</td>
<td>Noun Group (Acronym)</td>
</tr>
<tr>
<td></td>
<td>“The World Wide Web (WWW)...”</td>
<td></td>
</tr>
<tr>
<td>paper component</td>
<td>A component of the research paper. “...some successful applications (Section 3)...”</td>
<td>section, subsection</td>
</tr>
<tr>
<td>focus</td>
<td>The general focus. “A key focus of the technical specification was ...”</td>
<td>focus</td>
</tr>
</tbody>
</table>

Tab. 4.9: Some Concepts from the Conceptual Model

Relations refer to general activities of the author during the research and writing of the work: studying (investigate, study, etc.), reporting the work (present, report, etc.), motivating (objective, focus, etc.), thinking (interest, opinion, etc.), identifying (define, describe, etc.). Some of these relations are presented in Table 4.10 (the complete list is included in Appendix D).
<table>
<thead>
<tr>
<th>Relation</th>
<th>Explanation &amp; Example</th>
<th>Lexical Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>make known</td>
<td>Introducing the topic of the paper.</td>
<td>describe, present,</td>
</tr>
<tr>
<td></td>
<td>“In this paper we present...”</td>
<td>…</td>
</tr>
<tr>
<td>investigate</td>
<td>Investigating. The phase transition in binary constraint satisfaction problems, i.e., is investigated.</td>
<td>investigate, …</td>
</tr>
<tr>
<td>explain</td>
<td>Explaining.</td>
<td>discuss, explain,</td>
</tr>
<tr>
<td></td>
<td>“The accuracy of a prediction based on the expected number of solutions is discussed …”</td>
<td>…</td>
</tr>
<tr>
<td>describe</td>
<td>Describing.</td>
<td>compose, form, …</td>
</tr>
<tr>
<td>advantage</td>
<td>Identifying advantage. “... simulated annealing and evolutionary programming outperform back propagation.”</td>
<td>to have advantage</td>
</tr>
<tr>
<td>identify</td>
<td>Characterizing entity. “...a new algorithm called OPT-2 for optimal pruning...”</td>
<td>contain, classify,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>effective</td>
<td>Identifying effectiveness. “Our algorithm is effective for...”</td>
<td>to be effective, …</td>
</tr>
</tbody>
</table>

Tab. 4.10: Some Relations from the Conceptual Model

We have identified 52 types of information for the process of automatic text summarization referring to the following aspects of the technical article: background information (situation, need, problem, etc.), reporting of information (presenting entities, topic, sub-topics, objectives, etc.); referring to the work of the author (study, investigate, method, hypothesis, etc.); cognitive activities (argue, infer, conclude, etc.); and elaboration of the contents (definitions, advantages, etc.).

Concepts and relations are the basis for the classification of types of information referring to the essential contents of a technical abstract. Nevertheless, the single occurrence of a concept or relation in a sentence is not sufficient to understand the type of information it conveys.

The co-occurrence of concepts and relations in appropriate linguistic-conceptual patterns is used in our case as the basis for the classification of the sentences. Here we present only a few types of information (for the complete list refer to Appendix E):
4.4 FROM SOURCE TO ABSTRACT

**Topic of Document:** The author explicitly marks the topic of the document. This is identified in sentences from first or last sections of the document containing the make known relation, and concepts like author and research paper.
Ex.: In this paper we have presented a more efficient distributed algorithm which construct a breadth-first search tree in an asynchronous communication network.

**Author Development:** The explicit mention of a development of the author. We identify this information by the co-occurrence of the author concept and create relation.
Ex.: As part of the UK Electronic Libraries programme, the authors have developed a simple decision support tool which allows a library manager to compare the total cost of acquiring a given item of information from each of a number of different sources.

**Goal of Entity:** The explicit mention of the objective of a non conceptual entity. This is marked by the objective concept or relation.
Ex.: The goal of CCAD is to support exploratory design, while keeping the user central to the design activity.

**Description of Entity:** An entity is being described. This is identified by the describe relation.
Ex.: The algorithm is based on dynamic programming.

The types of information are classified in Indicative or Informative depending on the type of abstract they will contribute to. For example, Topic of Document and Author Development are indicative while Goal of Entity and Description of Entity are informative.

### 4.4 From Source to Abstract

According to Cremmins (1982), the last step in the production of the summary text is the “extracting” into “abstracting” step in which the extracted information will be mentally sorted into a pre-established format and will be “edited” using cognitive techniques. The editing of the raw material ranges from minor to major operations. However, he gives little indication about the process of editing. Major transformations are those of the complex process of language understanding and production such as deduction, generalization and paraphrase. Some examples of edition given by Cremmins are shown in Table 4.11.
<table>
<thead>
<tr>
<th>Edited Material</th>
<th>Source Material</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality in rats and mice of both sexes was dose related.</td>
<td>There were significant positive associations between the concentrations of the substance administered and mortality in rats and mice of both sexes.</td>
</tr>
<tr>
<td>No treatment related tumors were found in any of the animals.</td>
<td>There was no convincing evidence to indicate that endrin ingestion induced any of the different types of tumors which were found in the treated animals.</td>
</tr>
</tbody>
</table>

**Tab. 4.11: Text Editing in Human Abstracting**

In the first example, the concept *mortality in rats and mice of both sexes* is stated with the wording of the source document, instead the concept expressed through *the concentrations of the substance administered* is stated with the expression *dose*. In the second example, the relation between the *tumors* and *endrin ingestion* is expressed through the complex nominal *treated related tumors*.

In his rules for abstracting, Bernier (1985) states that redundancy, repetition, and circumlocutions are to be avoided. He gives linguistic expressions which could be safely removed from the extracted sentences or re expressed in order to gain conciseness these include expressions such as *It was concluded that X* to be replaced by *X* and *It appears that* to be replaced by *Apparently*. Also Mathis and Rush (1985), indicate that some transformations in the source material are allowed such as concatenation, truncation, phrase deletion, voice transformation, paraphrase, division and word deletion. Rowley (1982) mentions the inclusion of the lead or topical sentence, the use of active voice, and advocates for conciseness. But in fact, the issue of edition in text summarization (either manual or automatic) has been systematically neglected. In our work, we partially address this by enumerating some transformations frequently found in our corpus which are computationally implementable.

The transformations are always conceptual in nature and not textual (they do not operate on the string level), even if some of them seem to take the form of simple string deletion or substitution. The rephrasing transformations we have identified are:

1. **Syntactic Verb Transformation**: some verbs from the source document are re-expressed in the abstract, usually in order to make the style impersonal. The person, tense and voice of the original verb is changed. Also, verbs that are used to state the topic of the document are generally expressed in present tense (in active or passive voice), the same applies for verbs introducing the objective of the research paper or investigation (this is a pragmatic situation usually acknowledged in the domain of abstract writing: objectives are reported in present tense and results in past tense). In Table 4.12, we present two examples.
4.4. FROM SOURCE TO ABSTRACT

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addresses the issue of scalability of structure discovery using Subdue.</td>
<td>Finally, we address the issue of scalability of structure discovery using Subdue.</td>
</tr>
<tr>
<td>Presents a more efficient Distributed Breadth-First Search algorithm for an asynchronous communication network.</td>
<td>In this paper we have presented a more efficient distributed algorithm which construct a breadth-first search tree in an asynchronous communication network.</td>
</tr>
</tbody>
</table>

Tab. 4.12: Syntactic Verb Transformation

In the first example, the main verb of the sentence, *address*, is re-stated in the third person, simple present, active voice, without grammatical subject. In this example, also the structural marker *finally*, is deleted. In the second example, the main verb *have presented* is re-stated in the third person, simple present, active voice, without grammatical subject. Note that in this text-type, it would be incorrect to present the topic in past tense (i.e., *have presented*). Also the concepts *this paper* (*research paper*) and *we* (*author*) are not expressed in the abstract.

(2) **Lexical Verb Transformation**: the domain verb from the source information is changed and re-stated in the impersonal form. The examples are presented in Table 4.13.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>The changes required for other protocols in the TCP/IP suite to accommodate SIP are discussed, as are the mechanisms available to allow gradual transition of the Internet from IP to SIP.</td>
<td>It identifies the changes required for other protocols in the TCP/IP suite to accommodate SIP, and explains the mechanisms available to allow gradual transition of the Internet from IP to SIP.</td>
</tr>
<tr>
<td>Simple mechanisms for introducing hierarchy into interdomain routing system, making it practical to route a truly large Internet, are described.</td>
<td>This article details simple mechanisms for introducing hierarchy into interdomain routing system, making it practical to route a truly large Internet.</td>
</tr>
</tbody>
</table>

Tab. 4.13: Lexical Verb Transformation

In the first example, the nominals are stated as in the original, instead of the domain verbs *identify* and *explain* of the source document, the abstractor decided to employ the verb *discuss*. In the second example, the information is reported with the verb *describe* instead than with the verb *detail* used on the source document. Additionally, the concept *this article* (*research paper*) is deleted.

(3) **Verb Selection**: the topic or subtopic of the document is introduced by a domain verb. Let see the examples in Table 4.14.
The running of existing socket applications over SNA networks, which requires support for transparently masking the differences between TCP/IP and SNA from the applications, is described.

I define an intelligent agent as a self-contained system that undertakes context-sensitive decision making and task enforcement in an open (or semi-open) environment.

Tab. 4.14: Verb Selection

The first example shows the selection of the verb *describe* to introduce a topic of the document. Some additional transformations are required in this case such as the nominalisation of the verb *run* into *the running* and the use of a relative clause introduced by *which*. In the second example, instead of presenting the definition of the topic, the actual topic is presented with the verb *give*.

(4) **Conceptual Deletion**: domain concepts such as *research paper, author, paper component*, etc. are omitted in the abstract.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gives the basic definitions for relation algebras and their representation together with some properties of representations.</td>
<td><em>Section 2</em> gives the basic definitions for relation algebras and their representations together with some properties of representations.</td>
</tr>
<tr>
<td>Reports work comparing speech and keying in the context of a simulated command and control application.</td>
<td>In <em>this paper, we</em> report work comparing speech and keying in the context of a (simulated) command and control application.</td>
</tr>
</tbody>
</table>

Tab. 4.15: Conceptual Deletion

In the first example in Table 4.15, the expression which refers to the paper component *Section 2* is not reported in the abstract. In the second one, the concepts *research article (this paper)* and *author (we)* are not expressed in the abstract. The verb *report* is changed consequently.

(5) **Concept re-expression**: domain concepts such as *author, research paper, author related entity* are stated in impersonal such as in the first example in Table 4.16 where the concept *our algorithm* is expressed in the abstract with the impersonal form *the algorithm*. In the second example, the entity *our Genie system* is stated in the impersonal form *Genie*. The information of the abstract comes from two different parts of the document.
4.4. FROM SOURCE TO ABSTRACT

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyzes the complexity of the algorithm, and gives some examples of performance on typical networks.</td>
<td>We analyse the complexity of our algorithm, and give some examples of performance on typical networks.</td>
</tr>
<tr>
<td>Genie is a prototype software infrastructure which enables providers and consumers of satellite imagery data to interact.</td>
<td>Our Genie system, a prototype software infrastructure, demonstrates that a software agent-based approach is a powerful, flexible, paradigm for implementing such large-scale, autonomous, distributed systems.</td>
</tr>
<tr>
<td></td>
<td>Intelligent middle-ware will enable providers and consumers of satellite imagery to interact on a large scale.</td>
</tr>
</tbody>
</table>

**Tab. 4.16: Concept re-expression**

6) **Structural Deletion**: discourse markers (contrast, structuring, logical consequence, adding, etc.) such as first, next, finally, however, although are deleted. In Table 4.17, the discourse marker Indeed, is not kept in the abstract. Additionally, the jargon the “Net” is presented using the more appropriate form The Internet, and the expression in order to facilitate is replaced by to facilitate.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Internet has become a powerful tool to facilitate the analysis, sharing and distribution of information regarding both local and global environmental issues.</td>
<td>Indeed, the “Net” has become a powerful tool in order to facilitate the analysis, sharing, and distribution of information regarding both local and global environmental issues.</td>
</tr>
</tbody>
</table>

**Tab. 4.17: Structural Deletion**

7) **Clause Deletion**: one or more clauses (principal or complement) of the sentence are deleted as shown in the examples in Table 4.18 where only the complement phrase introduced by that is preserved.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>The designer description of an application contains much information that is useful in explaining its working.</td>
<td>The work described in this paper addresses these by noting that the designer description of an application contains much information that is useful in explaining its working.</td>
</tr>
<tr>
<td>MT is a metatheory of a mechanized object theory.</td>
<td>To emphasize this fact we say that MT is a metatheory of a mechanized object theory.</td>
</tr>
</tbody>
</table>

**Tab. 4.18: Clause Deletion**

8) **Parenthetical Deletion**: some parenthetical expressions are eliminated. In the first example in Table 4.19, the parenthetical expression is deleted as well as the initial clause *It*
will show how (only the complement phrase introduced by how is preserved). In the second example, the parenthetical expression (in this case an acronym) is not expressed in the abstract.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extending the designer’s description of the information processing system can allow for the construction of applications that are self-explanatory.</td>
<td>It will show how extending the designer’s description of the information processing system (<em>with a language that details how changes within the application occurs</em>) can allow for the construction of applications that are self-explanatory.</td>
</tr>
<tr>
<td>Archon provides a software framework that assist interaction between the subcomponents of a distributed AI application and a design methodology that helps structure these interactions.</td>
<td>The Archon (<em>architecture for cooperative heterogeneous on-line systems</em>) provides a software framework that assists interaction between the subcomponents of a distributed artificial intelligence application, and a design methodology that helps structure these interactions.</td>
</tr>
</tbody>
</table>

Tab. 4.19: Parenthetical Deletion

(9) **Acronym Expansion**: acronyms introduced for the first time are presented along with their expansions or only the expansion is presented. Two examples are given in Table 4.20.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>The focus is on a subset of networks and communications application programming interfaces.</td>
<td>The work focuses on a subset of networks and communications APIs.</td>
</tr>
<tr>
<td>The network protocols used are SNA (Systems Network Architecture) and TCP/IP (Transmission Control Protocol/Internet Protocol).</td>
<td>The network protocols used are SNA and TCP/IP.</td>
</tr>
</tbody>
</table>

Tab. 4.20: Acronym Expansion

In the first example, the acronym APIs is expanded. Additionally, the information is restated with the expression *the focus is on*. In the second one, the acronyms *SNA and TCP/IP* are expanded.

(10) **Abbreviation**: a shorter expression is used to refer to an entity. This could be an acronym or an anaphoric expression. The example shows the noun *the National Railway Museum* abbreviated with *the NRM* acronym. In addition, the verb *discuss* is chosen to present the actual subtopic.
4.4. FROM SOURCE TO ABSTRACT

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discusses the future of digital imaging at the NRM.</td>
<td>The future of digital imaging at the National Railway Museum.</td>
</tr>
</tbody>
</table>

**Tab. 4.21: Abbreviation**

(11) **Merge**: information from several parts of the source document are merged in a single sentence, this is the usual case when reporting entities stated in titles and captioning. In Table 4.22, two titles are merged in a single sentence and presented as subtopics with the verb *describe*.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protocol selection, address mapping, and connection management <em>are also described</em>.</td>
<td>Protocol Selection</td>
</tr>
<tr>
<td></td>
<td>Address Mapping and Connection Management</td>
</tr>
</tbody>
</table>

**Tab. 4.22: Merge**

(12) **Split**: information from one sentence of the source document is presented in separate sentences in the abstract, this could be because the sentence of the source contains two different types of information. In Table 4.23, the information about the development of the system and its advantages is presented in two sentences. But instead in the original appears in one complex sentence.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>A tesseral temporal reasoning system has been designed, based on tesseral addressing and using tesseral arithmetic.</td>
<td>This has resulted in a tesseral temporal reasoning system, based on tesseral addressing and using tesseral arithmetic, which offers the advantage that it is directly compatible with existing GIS technology.</td>
</tr>
</tbody>
</table>

**Tab. 4.23: Split**

(13) **Complex Reformulation**: a complex reformulation takes place: this could involve several cognitive processes such as generalization and paraphrase. In Table 4.24, the information from different parts including a title, which contains the complete description of the entity being introduced, is reformulated.
Professional Abstract | Source Document
---|---
SCULPTOR, a 3D intuitive interactive modeling tool, is being developed to create a design environment for architects based on virtual interaction tools, fast graphic libraries, and new approaches in Artificial Intelligence. | SCULPTOR - an intuitive 3D modeling tool

The motivation for our work is to invent a design environment for architects, based on the most recent hard- and software developments. These are mainly virtual reality (VR) interaction tools, fast graphic libraries, and new approaches in Artificial Intelligence.

Tab. 4.24: Complex Reformulation

(14) **Noun Transformations** : we also have observed other transformations such as nominalisations, generalization, restatement of complex nominals, deletion of complex nominals, expansion of complex nominals (different classes of aggregation) and change of case. Several examples illustrate these situations in Table 4.25.

<table>
<thead>
<tr>
<th>Professional Abstract</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>...business telecommunication prices in Europe</td>
<td>Business telecommunications prices (UK, Sweden, France, Austria, Germany, ...</td>
</tr>
<tr>
<td>...the University of Liverpool, UK...</td>
<td>the University of Liverpool</td>
</tr>
<tr>
<td>...Maxcess Library System...</td>
<td>Maxcess Library Systems, Inc. with Maxcess Library System</td>
</tr>
<tr>
<td>Experiment 1...</td>
<td>The 1st experiment</td>
</tr>
<tr>
<td>....regulation of cable TV in the UK...</td>
<td>UK : regulation of cable TV</td>
</tr>
<tr>
<td>The integration of speech and natural language processing</td>
<td>Integrating Speech and Natural Language Processing</td>
</tr>
<tr>
<td>The Austrian telecommunication infrastructure</td>
<td>The Austrian situation in the field of telecommunication infrastructure</td>
</tr>
</tbody>
</table>

Tab. 4.25: Noun Transformations

(15) **No Transformation** : the information is reported as in the source.

As shown in the examples, the “edition” of the source material involve several transformations at a time. The distribution of the 15 transformation in our corpus is shown in Table 4.26.
### 4.5 Summary

In this chapter, we have studied relations between abstracts and their source documents, this study was motivated by the need to answer to the question of content selection in text summarization (Spark Jones, 1993b). We have also addressed here another important research question: how is the information expressed in the abstract.

Our study was based on the manual construction of alignments between sentences of professional abstracts and elements of source documents. In order to obtain an appropriate coverage, abstracts from different secondary sources and source documents from different journals were used.

<table>
<thead>
<tr>
<th>Transformation</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic Verb Transformation</td>
<td>48</td>
<td>15%</td>
</tr>
<tr>
<td>Lexical Verb Transformation</td>
<td>13</td>
<td>4%</td>
</tr>
<tr>
<td>Verb Selection</td>
<td>70</td>
<td>21%</td>
</tr>
<tr>
<td>Conceptual Deletion</td>
<td>43</td>
<td>13%</td>
</tr>
<tr>
<td>Conceptual re-expression</td>
<td>4</td>
<td>1%</td>
</tr>
<tr>
<td>Structural Deletion</td>
<td>7</td>
<td>2%</td>
</tr>
<tr>
<td>Clause Deletion</td>
<td>47</td>
<td>14%</td>
</tr>
<tr>
<td>Prenominal Deletion</td>
<td>10</td>
<td>3%</td>
</tr>
<tr>
<td>Acronym Expansion</td>
<td>7</td>
<td>2%</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>3</td>
<td>1%</td>
</tr>
<tr>
<td>Merge</td>
<td>124</td>
<td>38%</td>
</tr>
<tr>
<td>Split</td>
<td>3</td>
<td>1%</td>
</tr>
<tr>
<td>Complex Reformulation</td>
<td>75</td>
<td>23%</td>
</tr>
<tr>
<td>Noun Transformation</td>
<td>70</td>
<td>21%</td>
</tr>
<tr>
<td>No Transformation</td>
<td>35</td>
<td>11%</td>
</tr>
</tbody>
</table>

**Tab. 4.26: Distribution of the 15 Transformation in the Corpus**

We found that transformations involving domain verbs appeared in 40% of the sentences, nouns editing occurred in 38% of the sentences, discourse level editing occurred in 19% of the sentences, merge and split of information occurred in 38% of the sentences, complex reformulation account for 23% of the sentences, and finally, only 11% of the information from the source document is stated without transformation.

While most approaches to automatic text summarization present the extracted information in both the order and the form of the original, this is not the case in human produced abstracts. Nevertheless, some transformations in the source document can be implemented by computers with state of the art techniques in natural language processing in order to improve the quality of the automatic summaries. This issue will be elaborated in the following chapter.
We have shown that:

1. More than 70% of the information for abstracts comes from introduction, conclusion, titles and captioning of the source document. This is an empirical verification of what is generally acknowledged in practical abstract writing in professional settings. In Chapter 5, Section 5.5, we show how the structural parts of the document are used in order to automatically compute the content of an abstract.

2. The information brought into the abstract has a well-defined status and can be identified on the basis of general concepts and relations referring to the technical article. This motivated the specification of a conceptual and linguistic model for text summarization. The implementation of this model in an automatic system is detailed in Chapter 5, and in particular we show through Sections 5.3 to 5.6 how the content of the abstract can be automatically identified and extracted.

3. We have identified 15 types of transformations usually applied in the source document in order to produce a coherent piece of text. 89% of the sentences of our corpus have been edited. The issue of text edition in text summarization has been until now practically ignored. In Chapter 5, Section 5.7 we detail the specification of patterns of sentence and text production inspired from our corpus study that were implemented in an automatic system.

Our initial corpus was expanded with 100 abstracts and source documents from the journal Computer & Control Abstracts, mainly in the field of Computer Science, in order to validate and enrich our model. While the linguistic information for our model has been manually collected, Teufel (1998) has shown how this labor-intensive task can be accomplished in a semi automatic fashion. The analysis presented here and the idea of the alignments was greatly influenced by the explorations of abstracting manuals (Cremmins, 1982) and work on cognitive science (Endres-Niggemeyer et al., 1991). Our conceptual model comes mainly from the empirical analysis of the corpus, but was also influenced by work on discourse modeling (Liddy, 1991) and in philosophy of science (Bunge, 1967). It is interesting to note that our concerns regarding the presentation and editing of the information for text summarization are now being addressed by other researchers as well. Jing and McKeown (2000) and Jing (2000) propose a cut-and-paste strategy as a computational process of automatic abstracting and a sentence reduction strategy in order to produce concise sentences. Knight and Marcu (2000) propose a noisy channel model and a decision-based model for sentence reduction also aiming at conciseness.
Chapitre 5

Selective Analysis

In this chapter, we present Selective Analysis, a method for text summarization of technical articles whose design is based on the study of the corpus described in the previous chapter. The method emphasizes the selection of particular types of information and its elaboration exploring the issue of dynamic summarization. We show how the information from the source document is selected and integrated in a new text using some heuristics and simple schemas of sentence and text production. While some editing is incorporated into the model, most of the information is presented as found in the source document.

5.1 Introduction

Our method is designed to produce short indicative and informative abstracts for technical articles. The function of the indicative abstract is to point to some of the main topics or themes considered in the source document (what the author presents, discusses, addresses, etc.) while the function of the informative abstract is to expand the main topics to provide the reader with specific types of information about them such as definitions, descriptions and statements of relevance.

This is done in two steps: first, the reader gets the indicative part of the abstract and a list of topics, which are terms obtained from the indicative abstract that are available for expansion; should the reader be interested in knowing more about the topics (s)he will get text spans from the source document elaborating them. While the indicative abstract depends on the structure, content, and to some extent, on specific types of information generally reported in this kind of summary, the informative abstract depends on the interest of the reader to know more about the topics. We have implemented this method in a computer system called SumUM\textsuperscript{1}, which will be described in Chapter 6.

The method is independent of any particular implementation. Nevertheless, it was designed keeping in mind the needs for accessing the content of long documents and the limitations of natural language processing of domain independent texts. The general architecture

\textsuperscript{1}Summarization at Université de Montréal
of the system is depicted in Figure 5.1.

![Diagram](image)

**Fig. 5.1: Main Processes in Selective Analysis**

The method consists of the following main steps:

- **Pre-processing and Interpretation**: its function is to identify the structure of the document and to interpret sentences according to our conceptual model (Chapter 4). This is described in Section 5.3;

- **Indicative Selection**: it aims to identify sentences of indicative type in order to construct the content for the indicative abstract. This is described in Section 5.5;

- **Informative Selection**: its function is to identify sentences of informative type in order to construct the content for the informative abstract. This is described in Section 5.6; and
5.2. THE INPUT

- **Indicative and Informative Generation**: for the presentation of the indicative abstract to the reader and the expansion of the indicative text with topic elaborations depending on the reader’s interests. These are described in Section 5.7.

5.2 The Input

The input to the system is a technical article in English without any mark-up containing the following structural elements:

- Title of the article;

- Author information and affiliation;

- Introductory section;

- Main sections without prototypical structure;

- References; and optionally

- Author abstract, list of keywords, and acknowledgements.

This specification of the input is to be considered as a precondition for the application of the method. An example of input text which will be used in the complete example provided in Section 5.8 is shown in Figure 5.2.
Features 3D scanning systems for rapid prototyping

Jim Clark
Jim Clark is an Industrial Applications Engineer with 3D Scanners Ltd...

Keywords: Motor industry, Rapid prototyping, Sensors, 3D

Type of Article: Technical

Discusses the use of two commercially available non-contact laser triangulation sensors....

Introduction

Rapid prototyping of industrial parts and the acquisition of 3D models are increasingly important stages in the mechanical manufacturing process. For example, in the automotive ...

REVERSA

REVERSA is a dual viewpoint non-contact laser scanner which comes complete with scanning software and data manipulation tools. The sensor head is shown in Figure 2 REVERSA sensor head.

ModelMaker

The ModelMaker scanning system (Figure 3 The Model Maker scanning system ) is a combination of a 3D laser stripe sensor, ...

Conclusions

Two non-contact scanning systems, REVERSA and ModelMaker, have been described and their application in industry demonstrated...

Fig. 5.2: Sample Input Text

5.3 Pre-processing and Interpretation

The purpose of this step is to analyze the input text in order to identify its structure (e.g., titles, author information, main sections and references) and interpret its content. This process is done using a segmentation algorithm developed for the purpose of this research and which is fully described in Section 6.2. This process is essential since the selection of information for an abstract is highly influenced by the documents’ structure, as we have shown in Chapter 4.
The main sections are syntactically and semantically interpreted in the following way: first, a text tagger is applied (details about the tagger are given in Section 6.2) which identifies the lexical category and citation form of each word. Then, in each section, titles and sentences are detected and interpreted using the information provided by the tagger. The interpretation consists of the application of a parser (cascade of finite state transducers) to identify some syntactic constructions such as noun, verb, adjectival and adverbial groups, and domain specific constructions (e.g., the parser identifies that the string “the following Section” is a reference to a paper component as soon as it scans it). The process associates semantics to each construction and interprets each element according to the conceptual model (this is fully described in Section 6.3). In order to do that, we use the conceptual dictionary (described in detail in Section 6.3) which associates lexical items with elements of the conceptual model: noun groups are interpreted as concepts, verb groups as relations and adjective groups as qualities (as we did in the specification of concepts and relations in Section 4.3). For example, the string “a straightforward process” is interpreted as the concept method because the word “process” is associated with that particular concept; the string “have been studied” is interpreted as the relation study because the verb “study” is associated with that relation in the conceptual model; the string “the most important aspects” is interpreted as a relevance quality of our conceptual model and will be used to identify relevant entities like in “The most important aspects of our system are”; and finally, the string “the algorithm” has no particular interpretation because the noun “algorithm” does not belong to any concept in our conceptual model. The actual implementation of SumUM does not deal with ambiguities, so different interpretations can be associated to the same syntactic construction.

SumUM represents each sentence as a list of lexical and syntactic elements enriched with semantic labels that indicate their interpretation. The representation is fully described and exemplified in Section 6.3.

SumUM uses the conceptual information it finds in the sentence to classify it according to types of information (the concepts, relations and types of information are given in Appendices C, D, and E). For example, if SumUM detects that a sentence contains the domain relation make known and a reference to a paper component, then it will classify this sentence as Topic of Section. In Tables 5.1 and 5.2, we present the information used by SumUM in order to classify sentences according to types of information. The sentence position in the source document and its type are recorded in the conceptual index (in the previous example, the sentence position and the type Topic of Section will be recorded). No ambiguities are resolved in SumUM.
<table>
<thead>
<tr>
<th>Indicative Type</th>
<th>Criteria for Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situation</td>
<td>(date (dc) \lor past perfect \lor past progressive \lor situation (dc/dr)) \land first section</td>
</tr>
<tr>
<td>problem/solution</td>
<td>(problem (dc/dr) \lor solution (dc/dr) \lor difficulty (dc) \lor difficult (da)) \land first section</td>
</tr>
<tr>
<td>Need</td>
<td>(need (dr) \lor necessity (dc) \lor necessary (da)) \land first section</td>
</tr>
<tr>
<td>entity introduction</td>
<td>identify (dr) \land first section</td>
</tr>
<tr>
<td>entity identification</td>
<td>identify (dr) \land \neg first section \land \neg last section</td>
</tr>
<tr>
<td>Topic</td>
<td>make known (dr) \land (author (dc) \lor research paper (dc) \lor study (dc) \lor research (dc) \lor work (dc) \lor institution (dc) \lor project (dc) \lor research group (dc))</td>
</tr>
<tr>
<td>Topic Description</td>
<td>make known (dr) \land (author (dc) \lor research paper (dc) \lor study (dc) \lor research (dc) \lor work (dc) \lor institution (dc) \lor project (dc) \lor research group (dc))</td>
</tr>
<tr>
<td>Possible Topic</td>
<td>(make known (dr) \land passive voice) \land (first section \lor last section)</td>
</tr>
<tr>
<td>Research Goal</td>
<td>objective (dr) \land (author (dc) \lor research paper (dc) \lor study (dc) \lor research (dc) \lor work (dc) \lor institution (dc) \lor project (dc) \lor research group (dc))</td>
</tr>
<tr>
<td>Conceptual Goal</td>
<td>conceptual objective (dc)</td>
</tr>
<tr>
<td>Focus</td>
<td>focus (dr) \land (author (dc) \lor research paper (dc) \lor study (dc) \lor research (dc) \lor work (dc) \lor institution (dc) \lor project (dc) \lor research group (dc))</td>
</tr>
<tr>
<td>Conceptual Focus</td>
<td>conceptual focus (dc)</td>
</tr>
<tr>
<td>Development</td>
<td>create (dr)</td>
</tr>
<tr>
<td>Interest</td>
<td>interest (dr) \land (author (dc) \lor research paper (dc) \lor study (dc) \lor research (dc) \lor work (dc) \lor institution (dc) \lor project (dc) \lor research group (dc))</td>
</tr>
<tr>
<td>Study</td>
<td>study (dr) \lor investigate (dr)</td>
</tr>
<tr>
<td>Method</td>
<td>method (dc)</td>
</tr>
<tr>
<td>Experiment</td>
<td>experiment (dc/dr)</td>
</tr>
<tr>
<td>Result</td>
<td>result (dc)</td>
</tr>
<tr>
<td>Inference</td>
<td>infer (dr)</td>
</tr>
<tr>
<td>knowledge</td>
<td>(explain (dr) \lor discover (dr)) \land (author (dc) \lor research paper (dc) \lor study (dc) \lor research (dc) \lor work (dc) )</td>
</tr>
<tr>
<td>Summarization</td>
<td>summarize (dc/dr)</td>
</tr>
<tr>
<td>Topic of Section</td>
<td>make known (dr) \land paper component (dc)</td>
</tr>
<tr>
<td>Signaling Structural</td>
<td>structural (dc)</td>
</tr>
<tr>
<td>Signaling Concept</td>
<td>analysis (dc) \lor application (dc) \lor comparison (dc) \lor concept (dc) \lor conclusion (dc) \lor definition (dc) \lor description (dc) \lor detail (dc) \lor development (dc) \lor discussion (dc) \lor explanation (dc) \lor introduction (dc) \lor method (dc) \lor outline (dc) \lor overview (dc) \lor presentation (dc) \lor recommendation (dc) \lor result (dc) \lor review (dc) \lor study (dc) \lor suggestion (dc) \lor summary (dc) \lor survey (dc)</td>
</tr>
</tbody>
</table>

Tab. 5.1: Information used to Classify Sentences in Indicative Types using Domain Concepts (dc), Domain Relations (dr) and Domain Adjectives (da)
<table>
<thead>
<tr>
<th>Informative Type</th>
<th>Criteria for Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>define (dr)</td>
</tr>
<tr>
<td>Description</td>
<td>describe (dr)</td>
</tr>
<tr>
<td>Elaboration</td>
<td>elaborate (dr)</td>
</tr>
<tr>
<td>Advantage</td>
<td>advantage (dc/dr)</td>
</tr>
<tr>
<td>Development</td>
<td>create (dr) ∨ development (dc)</td>
</tr>
<tr>
<td>Interest</td>
<td>interest (dr) ∨ interesting (da)</td>
</tr>
<tr>
<td>Relevance</td>
<td>relevance (da) ∨ essential (da)</td>
</tr>
<tr>
<td>Identification</td>
<td>identify (dr)</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>unique (da)</td>
</tr>
<tr>
<td>Study</td>
<td>study (dr) ∨ investigate (dr)</td>
</tr>
<tr>
<td>Usefulness</td>
<td>practical (da) ∨ use (dr) ∨ application (dc)</td>
</tr>
<tr>
<td>Goal</td>
<td>objective (dc/dr)</td>
</tr>
<tr>
<td>Focus</td>
<td>focus (dc/dr)</td>
</tr>
<tr>
<td>Positiveness</td>
<td>positive (da)</td>
</tr>
<tr>
<td>Novelty</td>
<td>novel (da)</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>effective (da)</td>
</tr>
<tr>
<td>Need</td>
<td>need (dr) ∨ necessity (dc) ∨ necessary (da)</td>
</tr>
<tr>
<td>Characteristics</td>
<td>characteristic (dc)</td>
</tr>
</tbody>
</table>

Tab. 5.2: Information used to Classify Sentences in Informative Types using Domain Concepts (dc), Domain Relations (dr) and Domain Adjectives (da)

The process then extracts terms and their semantics from the sentences. Terms are noun groups in citation form without determinants and their semantics is the citation form of the head noun (i.e., from the noun group “three on-line algorithms” the term “on-line algorithm” is obtained along with its semantics “algorithm”). **SumUM** gradually computes term and word frequencies and records the positions where the terms appeared. All this information about the terms is recorded in the term tree. **SumUM** also extracts acronyms and their expansions from the sentence and records that information in the acronym information structure (a simple algorithm that looks for patterns of acronyms and expansions was implemented for this task).

The final step is to produce the topical structure of the text which is the list of terms and words appearing on the title of the document and the titles of sections. Note that **SumUM** knows which elements are titles in order to extract the right terms. The information about the structure of the article is recorded in the text representation (sentences, titles, paragraph structure, etc.).
5.4 Representing the Information for the Abstract

We have decided to represent the information for both the indicative and the informative abstracts with templates. Templates are simple data structures composed of slot-filler pairs that record information for the tasks of content selection and text re-generation. The slot name in a template somehow indicates the semantics of the information of its filler which is a datum of specific type (numeric, alphanumeric, string, parsed sentence fragment, etc.).

In Chapter 2, we have presented one indicative abstract in Figure 2.5 and one informative abstract in Figure 2.6. We have observed that a sentence like “The work of Consumer Advise Centres is examined.” in the indicative abstract is used to introduce a topic (e.g., “The work of Consumer Advise Centres”) while a sentence like “CACs have dealt with pre-shopping advice, education on consumers’ rights and complaints about goods and services, advising the client and often obtaining expert assessments.” from the informative abstract gives actual information about that topic. In order to make evident this difference in the status of the information reported in a sentence, we work with two kinds of templates: indicative templates (Section 5.4.1); and informative templates (Section 5.4.2). A sentence will be used to instantiate a template of type T only if:

- it was classified as type T during interpretation;
- it matches a specific pattern associated with type T; and
- it satisfies some restrictions during the process of template instantiation.

The patterns used in the implementation of SumUM and the restrictions applied for instantiation were obtained during analysis of the corpus and testing of SumUM. In our approach patterns are sequences of concepts, relations, words, syntactic categories and variables $X_i$, for example: $X_1 + \text{experiment (dc)} + \text{Prep} + X_2 + \text{experiment (dr)} + X_3$ which matches a sequence of zero or more elements ($X_1$) followed by a concept experiment followed by a preposition followed by a relation experiment and a sequence of zero or more elements (the sentence “Some experiments on automatic abstracting were conducted at …” matches this pattern).

In this chapter we will give the complete specification of the templates and some examples of the process of instantiation (see Section 5.8) while the complete computational process of template instantiation is given in Chapter 6 (Sections 6.4 and 6.5).

5.4.1 Indicative Templates

The indicative templates contain the following mandatory slots:

- Type that specifies the type of the template (an alphanumeric constant),
- Id that contains a unique identifier (an integer),
- Position that contains the position of the sentence that was used to instantiate the template (section number and sentence number),
- **Topic candidates** filled in with a list of the terms mentioned in the sentence (not with all the terms from the sentence that is used for instantiation),
- **Weight** instantiated with an integer representing the “relevance” of the terms.

Term relevance is computed using the following formula:

$$\text{relevance}(\text{Term}) = \frac{\sum_{N \in \text{Term} \land \text{noun}(N)} \text{noun}_\text{frequency}(N)}{|N : N \in \text{Term} \land \text{noun}(N)|}$$

where \(\text{noun}(N)\) is true if \(N\) is a noun, \(\text{noun}_\text{frequency}(N)\) is a function computed during pre-processing and interpretation that gives the word count for noun \(N\) and the notation \(|S|\) stands for the number of elements in the set \(S\). As complex terms have lower distribution than single terms, this formula gives us an estimate of the distribution of the term and its components in the document. Note that in doing so, a term like “robot architecture” with low frequency in the whole document, will have great relevance because of the distribution of its component nouns “robot” and “architecture”.

The **Weight** slot is filled in with the following value:

$$\text{Template.Weight} = \sum_{\text{Term} \in \text{Template.Candidates}} \text{relevance}(\text{Term})$$

We now give the complete set of templates used in **SumUM**.

**Situation**: records information about the situation like in the sentence “Robotics approaches have been applied since the late 1970s with more and more advanced devices and strategies.” that contains two markers of situation: a concept **date** (e.g., “the late 1970s”) and a verb in the past perfect (e.g., “have been applied”). Its specification is given in Figure 5.3. The main slot is **Situation** containing the complete sentence without parenthetical or references.

<table>
<thead>
<tr>
<th>Type</th>
<th>situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Situation</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Situation filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

**Fig. 5.3**: Specification of the Situation Template

**Problem Identification**: records information from a sentence identifying a problem like in “Another problem is the workpieces are very large.” that contains the concept **problem** and matches the pattern: \(X_1 + \text{problem (dc)} + \text{define (dr)} + X_2\). The specification is given in Figure 5.4. The main slots are the **Marker** slot used to record an instance of the concept
problem (e.g., “Another problem”), the slot About that records the description of the problem (e.g., “the workpieces are very large”), and the Content slot that records the complete sentence.

<table>
<thead>
<tr>
<th>Type</th>
<th>problem identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Marker</td>
<td>instance of concept problem</td>
</tr>
<tr>
<td>About</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Content</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the About filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

**Fig. 5.4: Specification of the Problem Identification Template**

**Problem Solution**: records information from sentences containing references to problem/solution structures like in “Learning methods have been applied so far essentially to solving problems related to dealing with a static environment, e.g., learning the topography of an office environment, recognizing faces from sets of static images.” This sentence contains a solution relation and a concept problem and matches the pattern: \( X_1 + \text{solution} (dr) + \text{problem} (dc) + X_2 \). This template is specified in Figure 5.5: the main slots are the Problem slot that contains a description of the problem (e.g., “problems related to ...”), the Solution slot that contains a description of the solution (e.g., “Learning methods have been applied ...”) and the slot Content that contains the complete sentence.

<table>
<thead>
<tr>
<th>Type</th>
<th>problem solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Problem</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Solution</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Content</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Problem and Solution fillers</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

**Fig. 5.5: Specification of the Problem Solution Template**

**Need for Research**: records information about a need like in the sentence “Off-line programming systems require the availability of CAD data describing the workpieces to be welded.” This sentence contains a need relation (e.g., “require”) and matches a pattern: \( X_1 + \text{need} (dr) + X_2 \). The main slots are the Situation slot that records the entity in need (e.g., “Off-line programming systems”), the Necessity slot that records the actual need (e.g., “the availability of ...”) and the Content slot that records the complete sentence. The specification is shown in Figure 5.6.
5.4. REPRESENTING THE INFORMATION FOR THE ABSTRACT

<table>
<thead>
<tr>
<th>Type</th>
<th>need</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Situation</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Necessity</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Content</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Situation filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

**Fig. 5.6: Specification of the Need for Research Template**

**Entity Introduction**: records information about entities being introduced in the first section of the document like in the sentence “The Laboratory for Flexible Production Automation is partner in a cooperative project, named Delft Intelligent Assembly Cell (DIAC), between four faculties of Delft University of Technology, partly financed by SPIN, a funding by Dutch government.” This sentence matches the pattern of introduction: \( X_1 + \text{Prep} + X_2 + \text{identify (dr)} + X_3 \). Its specification is given in Figure 5.7: the main slots are the Entity slot that records the entity being introduced (e.g., “The Laboratory for Flexible Production Automation”) and the slot What that contains the information about the entity (e.g., “is partner in a cooperative project, ...”).

<table>
<thead>
<tr>
<th>Type</th>
<th>entity introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Entity</td>
<td>noun group</td>
</tr>
<tr>
<td>What</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Entity filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

**Fig. 5.7: Specification of the Entity Introduction Template**

**Topic of Document**: records information about the explicit introduction of the topic by the author like in the sentence “In this paper, we present details of our SWERS.” This sentence contains a research paper concept and a make known relation and matches the pattern of topic: \( X_1 + \text{author (dc)} + \text{make known (dr)} + X_2 \). The specification is given in Figure 5.8. The main slots are the Predicate slot that records information about the verb introducing the topic (e.g., “present”) and the What slot that records the information introduced by the verb (e.g., “details of our SWERS”).
Type: topic
Id: integer
Predicate: instance of make known relation
Where: instance of \{research paper, study, work, research\}
Who: instance of \{research paper, author, study, work, research\}
What: parsed sentence fragment
Position: section and sentence id
Topic candidates: list of terms from the What filler
Weight: number

Fig. 5.8: Specification of the Topic of Document Template

**Topic Description**: records information about the explicit description of the topic by the author like in the sentence "Another interesting activity, also reported here, was the realization, within the framework of a EUREKA project, of a tele-manipulator for servicing a new concept of urban infrastructures." that matches the pattern: $X_1 + \text{make known (dr)} + \text{research paper (dc)} + X_2$. The main slot is Description that records the description of the topic avoiding the elements that help identify the topic (e.g., "also reported here"). The specification is shown in Figure 5.9.

Type: topic description
Id: integer
Description: parsed sentence fragment
Position: section and sentence id
Topic candidates: list of terms from the Description filler
Weight: number

Fig. 5.9: Specification of the Topic Description Template

**Possible Topic**: records information from sentences introducing the topic implicitly (e.g., no concept research paper or author are present, but a relation make known in passive voice was detected) like in "An overview of different robotic systems developed for performing tasks in hazardous environments has been presented." that matches the pattern: $X_1 + \text{make known (dr)} + X_2$. The specification is presented in Figure 5.10. The main slots are the Predicate slot that records information about the verb introducing the topic (e.g., "has been presented"), the Argument slot that contains the main information introduced by the verb (e.g., "An overview of different ..."), and the slot Continuation containing additional information from the sentence.
5.4. REPRESENTING THE INFORMATION FOR THE ABSTRACT

<table>
<thead>
<tr>
<th>Type</th>
<th>possible topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Predicate</td>
<td>instance of make known relation</td>
</tr>
<tr>
<td>Argument</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Continuation</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Argument filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

Fig. 5.10: Specification of the Possible Topic Template

**Research Goal**: records information about the goal of the author or other entity like in the sentence “Because we aimed for a general robot control mechanism, as independent as possible from the particular environment and hardware used for the implementation, ANNs were more interesting to us than RL and EA techniques.” matching the pattern: $X_1 + \text{author (dc)} + \text{objective (dr)} + X_2$. The main slots are the Predicate slot that contains the verb (e.g., “aimed”), the Who slot that contains the conceptual entity (e.g., “we”), and the slot Goal that describes the goal (e.g., “for a general robot control ...”). The specification is shown in Figure 5.11.

<table>
<thead>
<tr>
<th>Type</th>
<th>research goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Predicate</td>
<td>instance of objective relation</td>
</tr>
<tr>
<td>Who</td>
<td>instance of {author, research, institution, work, research paper, project}</td>
</tr>
<tr>
<td>Goal</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Goal filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

Fig. 5.11: Specification of the Research Goal Template

**Conceptual Goal**: records information referring to the explicit goal of a conceptual entity like in the sentence “The aim of the project is to obtain a reliable, working demonstration-system of a flexible assembly cell that is capable of assembling industrial products.” that matches the pattern of goal: $X_1 + \text{conceptual objective (dc)} + \text{define (dr)} + X_2$. The template is given in Figure 5.12. The main slots are the Marker slot that records the conceptual goal (e.g., “The aim of the project”), the slot Predicate that records the predicate introducing the goal (e.g., “is”) and the slot Goal that records the goal (e.g., “to obtain a reliable, ...”).
<table>
<thead>
<tr>
<th>Type</th>
<th>conceptual goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Marker</td>
<td>instance of {goal of paper, author, study, work, research, institution, project, group}</td>
</tr>
<tr>
<td>Predicate</td>
<td>instance of define relation</td>
</tr>
<tr>
<td>Goal</td>
<td>parsed fragment sentence describing the goal</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Goal filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

Fig. 5.12: Specification of the Conceptual Goal Template

**Focus** records information from sentences referring to the focus of a conceptual entity like in the sentence “After some successful developments in the field of industrial robotics, the department focused its attention on the area of robots for hostile/hazardous environments.” that matches the pattern: $X_1 + \text{institution (dc)} + \text{focus (dr)} + X_2$. The specification is presented in Figure 5.13. The main slots are the **Predicate** slot to record information about the verb introducing the focus of attention (e.g., “focused”), the **Marker** slot that records the conceptual entity (e.g., “the department”) and the **Focus** slot that records the focus (e.g., “the area of robots for ...”).

<table>
<thead>
<tr>
<th>Type</th>
<th>focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Predicate</td>
<td>instance of focus relation</td>
</tr>
<tr>
<td>Marker</td>
<td>instance of {author, paper, project, study, research, institution}</td>
</tr>
<tr>
<td>Focus</td>
<td>parsed fragment describing the focus</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Focus filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

Fig. 5.13: Specification of the Focus Template

**Conceptual Focus** records information referring to the explicit focus of a conceptual entity like in the sentence “The focus of this paper is to consider usability issues in catalogs independently of current trends on the Internet.” that matches the pattern: $X_1 + \text{conceptual focus (dc)} + \text{define (dr)} + X_2$. This template is presented in Figure 5.14. The main slots are the **Marker** slot to record the conceptual focus (e.g., “The focus of this paper”), the **Predicate** slot to record the verb introducing the focus (e.g., “is”), and the slot **Focus** to record the focus (e.g., “to consider ...”).
5.4. REPRESENTING THE INFORMATION FOR THE ABSTRACT

<table>
<thead>
<tr>
<th>Type</th>
<th>conceptual focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Marker</td>
<td>instance of {focus of paper, author, study, work, research, institution, project, group}</td>
</tr>
<tr>
<td>Predicate</td>
<td>instance of define relation</td>
</tr>
<tr>
<td>Focus</td>
<td>parsed fragment sentence describing the focus</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Focus filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

Fig. 5.14: Specification of the Conceptual Focus Template

Author Development: records the explicit mention of a development by the author or other conceptual entity like in “We implemented the DRAMA architecture in a number of experiments with wheeled LEGO or FISCHERTECHNIK robots which are widely used tools for research on mobile robots.” that matches the pattern: \( X_1 + \text{author (dc)} + \text{create (dr)} + X_2 \). The template is specified in Figure 5.15. Its main slots are the Predicate slot that records the verb referring to the development (e.g., “implemented”), the slot Who that contains the concept (e.g., “We”), and the slot Argument that contains the development (e.g., “the DRAMA architecture in ...”).

<table>
<thead>
<tr>
<th>Type</th>
<th>author development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Predicate</td>
<td>instance of create relation</td>
</tr>
<tr>
<td>Who</td>
<td>instance of {author, research, institution}</td>
</tr>
<tr>
<td>Argument</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Argument filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

Fig. 5.15: Specification of the Author Development Template

Something Developed: records information about sentences containing references to developments like in “The gantry was designed to accommodate the 12m x 12m panel size requirements in Keppel FELS, a shipyard in Singapore.” that matches the pattern of development: \( X_1 + \text{create (dr)} + X_2 \). The template specification is given in Figure 5.16. The main slots are the Predicate slot that records the information about the verb referring to the development (e.g., “was designed”), the Argument slot used to record the information about the thing developed (e.g., “The gantry”), and the Continuation slot that records additional information from the sentence (e.g., “to accommodate the ...”).
**Author Interest**: records information about sentences referring to the explicit mention of the interest of a conceptual entity like in “*For SWERS, we are interested in providing the robot with a walk-through programming capability.*” that matches the pattern: $X_2$+author (dc)+interest (dr)+$X_2$. This template is specified in Figure 5.17. Its main slots are the **Who** slot that contains the entity (e.g., “we”) and the slot **What** that records information about the interest (e.g., “*in providing the robot ...*”).

<table>
<thead>
<tr>
<th>Type</th>
<th>author interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Predicate</td>
<td>instance of interest relation</td>
</tr>
<tr>
<td><strong>Who</strong></td>
<td>instance of {research paper, author, study, work, research, group, institution, project}</td>
</tr>
<tr>
<td><strong>What</strong></td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td><strong>Position</strong></td>
<td>section and sentence id</td>
</tr>
<tr>
<td><strong>Topic candidates</strong></td>
<td>list of terms from the <strong>What</strong> filler</td>
</tr>
<tr>
<td><strong>Weight</strong></td>
<td>number</td>
</tr>
</tbody>
</table>

**Author Study**: records information about the explicit mention of the study of the author or other conceptual entity like in “*We have measured the cost to inspect the configuration files for all places.*” that matches the pattern: $X_1$+author (dc)+study (dr)+$X_2$. The template is given in Figure 5.18. The main slots are the **Predicate** slot that records the verb that introduces the study (e.g., “*have measured*”), the **Argument** slot that contains the thing studied (e.g., “*the cost*”), and the **Who** slot that records the entity (e.g., “*We*”).

<table>
<thead>
<tr>
<th>Type</th>
<th>development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td><strong>Predicate</strong></td>
<td>instance of create relation</td>
</tr>
<tr>
<td><strong>Argument</strong></td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td><strong>Continuation</strong></td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td><strong>Position</strong></td>
<td>section and sentence id</td>
</tr>
<tr>
<td><strong>Topic candidates</strong></td>
<td>list of terms from the <strong>Argument</strong> filler</td>
</tr>
<tr>
<td><strong>Weight</strong></td>
<td>number</td>
</tr>
</tbody>
</table>
5.4. REPRESENTING THE INFORMATION FOR THE ABSTRACT

<table>
<thead>
<tr>
<th>Type</th>
<th>author study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Predicate</td>
<td>instance of study relation</td>
</tr>
<tr>
<td>Who</td>
<td>instance of {research paper, author, study, work, research}</td>
</tr>
<tr>
<td>Argument</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Argument filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

**Fig. 5.18:** Specification of the Author Study Template

**Something Studied:** records information about a study like in “The time to complete parabolic trajectories between picking positions has been evaluated to be about 1.7 s (for a distance of 400mm between picking positions and a backward displacement of 250mm which in total represents a travelling length of about 850mm).” that matches the pattern of study: \( X_1 + \text{study (dr)} + X_2 \). The specification is given in Figure 5.19: its main slots are the Predicate slot that records the verb (e.g., “has been evaluated”), the Argument slot that records the thing studied (e.g., “The time to complete the parabolic trajectories ...”), and the Continuation slot to record additional information.

<table>
<thead>
<tr>
<th>Type</th>
<th>study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Predicate</td>
<td>instance of study relation</td>
</tr>
<tr>
<td>What</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Continuation</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the What filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

**Fig. 5.19:** Specification of the Study Template

**Method:** records information about the explicit mention of a method like in the sentence “Welding of the stiffeners vertically onto the plates is a straightforward process using current automated welding systems.” that matches the pattern: \( X_1 + \text{method (dc)} + \text{define (dr)} + X_2 \). The template is shown in Figure 5.20. The main slots here are the Marker slot used to record the concept method (e.g., “a straightforward process”), the Method slot to record the fragment naming or describing the method (e.g., “Welding of the ...”), and the slot Content that records the sentence without parenthetical and references.
**Experiment**: records information about the experiments like in the sentence “In these experiments the robot learned to recognise landmarks (boxes, light, aluminium covered ground areas).” that matches the pattern of experiment: $X_1 + \text{experiment (dc)} + X_2$. The template specification is given in Figure 5.21. Its main slots are the **Marker** slot that records the concept (e.g., “these experiments”), the slot **Experiment** that contains the description of the experiment (e.g., “the robot learned to ...”), and the slot **Content** that records the original sentence without parenthetical.

**Result**: records information about sentences referring to results like in “Results showed complete success for all teachings.” that matches the pattern: $X_1 + \text{result (dc)} + X_2$. The template is specified as in Figure 5.22: the main slots are the **Marker** slot that contains a reference to the concept **result** (e.g., “Results”), the slot **Result** that contains the fragment describing the result (e.g., in this case empty) and the slot **Content** containing the original sentence.
### 5.4. Representing the Information for the Abstract

<table>
<thead>
<tr>
<th>Type</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Marker</td>
<td>instance of the concept result</td>
</tr>
<tr>
<td>Result</td>
<td>noun group</td>
</tr>
<tr>
<td>Content</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from both Marker and Result fillers</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

**Fig. 5.22**: Specification of the Result Template

**Inference**: records information about sentences referring to an inference like in “It is concluded that the dynamic Domain Name System and the directory server look-up are the two best approaches for resolving dynamic IP addressing.” that matches the pattern: “It”+infer (dr)+X. This template is presented in Figure 5.23. The main slots are the **Main** slot to record the main inference (e.g., “that the dynamic Domain Name System ...”), and the **Inference** slot that records the sentence.

<table>
<thead>
<tr>
<th>Type</th>
<th>inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Main</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Inference</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the Main filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

**Fig. 5.23**: Specification of the Inference Template

**Author Knowledge**: records information about what the author (or other conceptual entity) found like in “We found nonlinear subscript expressions (some generated by our transformations and the rest part of the original program) in the Perfect Benchmarks.” that matches the pattern: $X_1$+author (dc)+discover (dr)+$X_2$. The template is specified in Figure 5.24. The main slots are the **Predicate** slot that contains the verb (e.g., “found”), the **Who** slot that contains the conceptual entity (e.g., “We”), and the **What** that contains the description of what was found (e.g., “nonlinear subscript ...”).
### Type: Knowledge
- **Id:** integer
- **Predicate:** instance of discover relation
- **Who:** instance of \{research paper, author, study, work, research\}
- **What:** parsed sentence fragment
- **Position:** section and sentence id
- **Topic candidates:** list of terms from the What filler
- **Weight:** number

#### Figure 5.24: Specification of the Knowledge Template

**Summarizing:** records sentences that identify a summary like in “To sum up, the method can be used in a variety of situations.” that matches the pattern: `summarize (dr)+X`. The template is shown in Figure 5.25. Its main slot is **Content** that record part of the sentence (e.g., “the method can be used ...”).

### Type: Summarization
- **Id:** integer
- **Content:** parsed fragment sentence d
- **Position:** section and sentence id
- **Topic candidates:** list of terms from the Content filler
- **Weight:** number

#### Figure 5.25: Specification of Summarization Template

**Entity Identification:** records sentences identifying entities like in “Certification includes extensive laboratory tests of welded specimens.” that matches the pattern: `X1+identify (dr)+X2`. The template is specified in Figure 5.26. Its main slots are the **Entity** slot that records the entity being identified (e.g., “Certification”) and the **What** slot containing the fragment identifying the entity (e.g., “includes extensive...”).

### Type: Entity Identification
- **Id:** integer
- **Entity:** noun group
- **What:** parsed sentence fragment identifying the entity
- **Position:** section and sentence id
- **Topic candidates:** list of terms from the Entity filler
- **Weight:** number

#### Figure 5.26: Specification of Entity Identification Template

**Topic of Section:** records information from sentences explicitly referring to the topic of a section like in “The fifth Section presents the benefits and conclusions.” that matches
the pattern: \[X_1 + \text{paper component (dc)} + \text{make known (dr)} + X_2.\] The template is given in Figure 5.27. The main slots are the **Predicate** slot that records the verb used to introduce the topic of the section (e.g., “presents”), the slot **Section** that records the number of the section (e.g., section(5)), and the **Argument** slot that records the topic being introduced (e.g., “the benefits and conclusions”).

<table>
<thead>
<tr>
<th>Type</th>
<th>topic of section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Predicate</td>
<td>instance of make known relation</td>
</tr>
<tr>
<td>Section</td>
<td>instance of section((\text{Id}))</td>
</tr>
<tr>
<td>Argument</td>
<td>parsed sentence fragment</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the <strong>Argument</strong> filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

**Fig. 5.27:** Specification of the Topic of Section Template

**Signaling Structural** : records information presented in structural elements like in the sentence “*Table I Important welding process parameters shows the fields in the welding database.*” that matches the pattern: \[X_1 + \text{structural (dc)} + X_2 + \text{show (dr)} + X_3.\] The template is shown in Figure 5.28. The main slots are the **Predicate** slot that records the verb used to present the information (e.g., “shows”) and the slot **Argument** that records the information referred to by the verb (e.g., “the fields in the welding database”).

<table>
<thead>
<tr>
<th>Type</th>
<th>signaling structural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Structure</td>
<td>instance of table((\text{Id})), figure((\text{Id})), picture((\text{Id})), plate((\text{Id}))</td>
</tr>
<tr>
<td>Predicate</td>
<td>instance of show structural relation</td>
</tr>
<tr>
<td>Argument</td>
<td>parsed sentence fragment describing the structural element</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>list of terms from the <strong>Argument</strong> filler</td>
</tr>
<tr>
<td>Weight</td>
<td>number</td>
</tr>
</tbody>
</table>

**Fig. 5.28:** Specification of Signaling Structural Template

**Signaling Concept** : records information about specific concepts referred to in the article like in the title of section “*Walk-through teaching method*” that matches the pattern: method (dc). The template is given in Figure 5.29: its main slots are the **Predicate** slot that contains a verb to be used in order to introduce the concept in the text of the abstract and it depends on the concept (e.g., “cover”) and the slot **What** that contains the concept (e.g., “Walk-through teaching method”).
Two special types of templates the Multi and Merged templates, which organize information of the same type together facilitating the generation of complex sentences, are specified in Figure 5.30. These are instantiated once SumUM has decided which information goes in the indicative abstract and so they are not used for content selection. This process is exemplified later in Section 5.7.

![Figure 5.30: Specification of Multi and Merged Templates](image)

### 5.4.2 Informative Templates

Informative templates are used to record information about specific topics that are to be presented only upon reader’s request. All informative templates are similar in structure and contain the following mandatory slots:

- **Type** that specifies the type of the template (an alphanumeric constant),
- **Id** that contains a unique identifier (an integer),
- **Topic** that contains a topic to be introduced in the abstract (a term),
- **Predicate** that contains a sentence fragment that was used to identify the information about the topic,
- **Position** that contains the position of the sentence (section and sentence position) used to instantiate the template, and
- **Content** that is instantiated with the particular information about the topic (generally the whole sentence).

In the actual implementation, SumUM instantiates the Content slot with the complete sentence elaborating the topic. The specification of informative templates is given in Figure 5.31. The types of informative templates are given below along with examples of instantiation of the slots Topic and Predicate.

![Figure 5.31: Specifying Informative Templates](image)
5.4. REPRESENTING THE INFORMATION FOR THE ABSTRACT

<table>
<thead>
<tr>
<th>Type</th>
<th>Informative Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>integer</td>
</tr>
<tr>
<td>Topic</td>
<td>term</td>
</tr>
<tr>
<td>Predicate</td>
<td>parsed fragment sentence</td>
</tr>
<tr>
<td>Content</td>
<td>parsed fragment sentence</td>
</tr>
<tr>
<td>Position</td>
<td>section and sentence id</td>
</tr>
</tbody>
</table>

**Figure 5.31: Specification of Informative Templates**

**Definition:** This template is intended to record information from sentences containing definitions of topical entities like in "IMA is a two-level software architecture for rapidly integrating these elements, for an intelligent machine such as a service robot." that matches the pattern: \( X_1 + \text{Topic} + \text{define} (dr) + X_2 \). The slot Topic is instantiated with "IMA" and the slot Predicate is instantiated with "is".

**Description:** Its function is to record information from sentences containing descriptions of topical entities like in the sentence "The conceptual design of a Web application consists of four components: a metaphor, a storyboard, a concept of user interaction, and an overall look and feel for the application." that matches the pattern: \( X_1 + \text{Prep} + \text{Topic} + \text{describe} (dr) + X_2 \). The slot Topic is instantiated with "Web application" and the Predicate slot with "consist".

**Elaboration:** Its function is to record information from sentences containing elaborations of topical entities like in the sentence "In fact, MAMAS provides a user-friendly graphical interface to operate directly on the system." that matches the pattern: \( X_1 + \text{Topic} + \text{elaborate} (dr) + X_2 \). The slot Topic is instantiated with "MAMAS" while the slot Predicate is instantiated with "provides".

**Advantage:** It records information from sentences containing references to the advantages of a topical entity like in the sentence "While the advantages of shell languages are mainly concentrated on the possibility of rapid application development, ..." that matches the pattern of advantage: \( X_1 + \text{advantage} (dc) + \text{Prep} + \text{Topic} + X_2 \). The slot Topic is instantiated with "shell languages" while the slot Predicate is instantiated with "the advantages".

**Development:** It is used to record information about the development of a topical entity like in "We have been developing robotic aid systems for the disabled." that matches the pattern: \( X_1 + \text{create} (dr) + \text{Topic} + X_2 \). The slot Topic is instantiated with "robotic aid systems" and the slot Predicate is instantiated with "have been developing".

**Interest:** It is used to record information about the interest of a topical entity like in the sentence "The user interface objectives address factors that affect the user and include ..." that matches the pattern: \( X_1 + \text{Topic} + \text{interest} (dr) + X_1 \). The slot Topic is instantiated with "The user interface objectives" and the slot Predicate with "address".
Relevance: It records information from sentences referring to the relevance of a topic like in the sentence “The most important information about a particular certificate - apart from the public key - is its validity.” that matches the pattern: $X_1$+relevance (da)+Prep+Topic+$X_2$. The Topic slot is instantiated with “a particular certificate” and the slot Predicate with “The most important information”.

Identification: It records information from sentences identifying the topic like in “Possible business objectives include: provide corporate image enhancement, improve customer service, attract prospective employees, market products, increase employee productivity, or provide an easy cross-platform approach to gathering and disseminating corporate data.” that matches the pattern: Topic+identify (dr)+$X$. The Topic slot is instantiated with the term “Possible business objectives” and the slot Predicate with “include”.

Uniqueness: Its function is to record information from sentences referring to the unique aspects of a topical entity like in “One of the unique features of the OLPS is the integrated approach to the planning of the welding operations.” that matches the pattern $X_1$+unique (da)+Prep+Topic+$X_2$. The slot Topic is instantiated with “the OLPS” and the slot Predicate with “the unique features”.

Study: It records information about the study of the topic like in “Most studies evaluate a motif only on binding peptides and do not test the ability of the motif to identify non-binding peptides.” that matches the pattern: $X_1$+study (dr)+Topic+$X_2$. In this case the slot Topic is instantiated with “a motif” and the slot Predicate with “evaluate”.

Usefulness: It records information about the practicality of a topic like in “While the models developed for NTRS are generally applicable, the current scope is limited to technical publications.” that matches the pattern $X_1$+Topic+define (dr)+practical (da)+$X_2$. The Topic slot is instantiated with “NTRS” and the slot Predicate with the fragment “are generally applicable”.

Goal: It records information about the objectives of a topical entity like in “The goal of NTRS is to provide one-stop-shopping for NASA technical publications.” that matches the pattern $X_1$+objective (dc)+Prep+Topic+$X_2$. The slot Topic is instantiated with “NTRS” and the slot Predicate with “The goal”.

Focus: It records information about the focus of a topical entity like in “These systems do not compete with the components of NTRS because NTRS is focussed only on the customer side of searching and retrieval.” that matches the pattern $X_1$+Topic+focus (dr)+$X_2$. Here, the slot Topic is instantiated with the term “NTRS” and the slot Predicate with “is focussed”.

Positiveness: Its function is to record information about the positive aspects of a topical entity like in the sentence “These enhancements over free WAIS make free WAIS-sf the best public-domain, general purpose, full-text indexing and search engine available today.” that
matches the pattern $X_1 + \text{Topic} + \text{positive} \ (da) + X_2$. The \text{Topic} slot is instantiated with the term “free WAIS-sf” and the slot \text{Predicate} with the fragment “the best public-domain”.

**Novelty** : It records information from sentences referring to the new aspects of a topical entity like in the sentence “To meet this goal, researchers at various NASA installations have developed several new methods of distributing information to the nation’s research and industrial sectors.” that matches the pattern $X_1 + \text{novel} \ (da) + \text{Prep} + \text{Topic} + X_2$. The slot \text{Topic} is instantiated with “distributing information” and the slot \text{Predicate} with the fragment “several new methods”.

**Effectiveness** : It records information from sentences referring to the goodness of a topical entity like in “In particular, free WAIS-sf16 is superior to the CNIDR free WAIS version in that it introduces the concepts of structured fields in the document that can be searched separately.” that matches the pattern $X_1 + \text{Topic} + \text{define} \ (dr) + \text{effective} \ (da) + X_2$. The slot \text{Topic} is filled in with the term “free WAIS-sf16” and the slot \text{Predicate} with the fragment “is superior”.

**Need** : It records information about the necessity of a topical entity like in “Top-down development of Web applications requires some understanding of the technology, the application area, and the creative possibilities in applying the technology to the problems in the application domain.” that matches the pattern $X_1 + \text{Topic} + \text{need} \ (dr) + X_2$. The slot \text{Topic} is instantiated with the term “Web applications” and the slot \text{Predicate} with the fragment “requires”.

**Characteristics** : Its function is to record information from sentences mentioning the characteristics of a topical entity like in “What are the design characteristics of computer systems that support distributed product development?” that matches the pattern $X_1 + \text{characteristic} \ (dc) + \text{Prep} + \text{Topic} + X_2$. The slot \text{Topic} is instantiated with the term “computer systems” and the \text{Predicate} slot with “the design characteristics”.

### 5.5 Indicative Selection

The purpose of this step is to identify potential topics of the document and to compute the content of the indicative abstract, i.e., a set of indicative templates containing information extracted from the text which introduces the topics. This is done by looking for indicative types of information through all the text using the conceptual index computed during pre-processing and interpretation. SumUM analyzes each sentence classified as indicative and if possible, instantiates indicative templates with information from the sentence (structural elements, references and parenthesized expressions are not extracted). In SumUM, this analysis looks for patterns given in the previous section to extract information. For example, a sentence like “Section 2 describes HuDL in greater detail and section 3 discusses system integration and the IMA.” is classified by SumUM as \text{Topic of Section} during interpretation because of the co-occurrence of the concept \text{paper component} (e.g., “Section 2”) and the relation \text{make known} (e.g., “describes”). This sentence is a candidate for
instantiation of a template of type **Topic of Section**, but in order to do so, **SumUM** first verifies if the sentence satisfies a pattern associated with the type **Topic of Section**. The sentence matches a pattern of this type which asks for a sequence of the following elements: a *paper component* concept ($X_1$) followed by a *make known* relation in active voice ($X_2$) followed by a non empty sequence of elements ($X_3$) followed by a conjunction and a *paper component*. The elements $X_i$ are used to instantiate a new template of type **Topic of Section** as follows: $X_1$ is used to instantiate the slot **Section**, $X_2$ is used to instantiate the slot **Predicate**, and $X_3$ is used to instantiate the slot **Argument**. The Id slot is instantiated with a fresh integer identifier and the **Topic Candidates**, **Position** and **Weight** slots are filled in as dictated by the specification. The instantiated template is shown later in Section 5.7 (Figure 5.32).

As only part of the information is used to instantiate the templates, transformations like **Split**, **Conceptual Deletion**, **Clause Deletion** described in Section 4.4 are done during this step. In the previous example, the sentence was split in order to fill a template of type **Topic of Section** and the concept *paper component* will not be expressed in the abstract implementing the **Conceptual Deletion** (refer to Section 5.7).

When instantiating templates of type **Signaling Structural** or **Signaling Concept**, an appropriate verb is chosen to instantiate the **Predicate** slot (**show** for information from figures; **present** for information from tables and concepts like **analysis** and **summary**; **overview** for the concept **overview**; and **cover** for concepts like **development** and **discussion**) implementing the **Select Verb** transformation through this process (refer to template **Signaling Concept** in Section 5.4.1).

The set of instantiated templates produced by the indicative selection is called **Indicative Data Base** (IDB).

The content of the indicative abstract is computed by the application of the following steps:

1. **SumUM** matches the terms of the **topical structure** with the terms appearing in the **topic candidates** slot in the templates in the IDB. Two terms $Term_1$ and $Term_2$ match if $Term_1$ is substring of $Term_2$ or if $Term_2$ is substring of $Term_1$ (i.e., **scanning** system matches **system**).

2. For each term in the **topical structure**, **SumUM** selects one template matching the term (if any): the one with the greatest **Weight**. If ties occur, **SumUM** selects a template in the following order (using the **Type** slot): **Topic of Document** > **Topic of Section** > **Topic Description** > **Possible Topic** > **Author Study** > **Author Development** > **Author Interest** > **Conceptual Goal** = **Research Goal** > **Conceptual Focus** = **Focus** > **Entity Introduction** > **Entity Identification** > **Signaling Structural** = **Signaling Concepts** > the other indicative types. This order gives preference to explicit topical information more usually found in indicative abstracts. If ties, the **Position** and the **Id** slots are used to solve the conflict: i.e., if
two **Topic** templates have the same **Weight**, the template with position closer to the beginning of the document is selected and if ties, the template with lower **Id** is used.

The instantiated templates selected by this process constitute the set **indicative content** of the abstract and the terms and words from the **Topic candidates** slots and their expansions constitute the set **potential topics** of the document. The expansion of a term $T$ is obtained in the **acronym information** (the expression acronym(A,Expansion) means that A is the acronym of Expansion) and by extracting from the **term tree** all the terms with the semantics of $T$ (i.e., $\text{term}_2 \in \text{expansion}(\text{term}_1)$ if $\text{term}_2 \in \text{term.tree} \land \text{semantics}(\text{term}_1) = \text{semantics}(\text{term}_2)$, the semantics of each term and the acronym information was computed during pre-processing and interpretation).

More formally, the **potential topics** are computed as follows:

$$\text{Candidates} = \bigcup_{T \in \text{indicative content}} T.\text{Topic.Candidates}$$

$$\text{Acronyms} = \{A : \exists \text{Term} \in \text{Candidates} \land \text{acronym}(A,\text{Term}) \lor \text{acronym}(\text{Term},A)\}$$

$$\text{Words} = \{\text{Word} : \exists \text{Term} : (\text{Term} \in \text{Candidates} \cup \text{Acronyms}) \land (\text{Word} \in \text{Term})\}$$

$$\text{Term Candidates} = \text{Candidates} \cup \text{Words} \cup \text{Acronyms}$$

$$\text{potential topics} = \bigcup_{T \in \text{Term Candidates}} \text{expansion}(T)$$

For a complete example refer to Section 5.8.

## 5.6 Informative Selection

This process determines the actual topics from the **potential topics** computed by the indicative selection and computes the content of the informative abstract (i.e., a set of instantiated informative templates). **SumUM** considers as topics for presentation to the reader only those terms that are elaborated in the source document in at least one of the informative types described in Section 5.4.2. In order to accomplish that objective it considers each term in the list **potential topic** at a time. For each potential topic $T$ and sentence $S$ where $T$ appears (that information is found on the **conceptual index**), **SumUM** verifies if $S$ contains an informative marker (see Section 4.3 and Table 5.2) and if it follows an informative pattern. If so, $S$ is used to instantiate a template $Tmpl$ of the appropriate type (Section 5.4.2) that is included in the **informative database**. The term $T$ is considered a topic of the document and will be included in the list **topics** computed using the following formula:

$$\text{topics} = \{\text{Term} : \exists \text{Template} : (\text{Template} \in \text{Informative Data Base}) \land \text{Term} = \text{Template.Topic}\}$$
An example about this step is given in Section 5.8 and the computational process implementing this step is described in Section 6.5. Potential topics without “elaborations” are not considered as topics by SumUM\textsuperscript{2}.

5.7 Generation

The process of generation consists of the arrangement of the information in a pre-established conceptual order, the merge of some types of information, and the reformulation of the information in one text paragraph.

**Conceptual Sort**: The indicative content is sorted using positional information and the following “conceptual” order:

1. templates of type **Problem Solution**, **Problem Identification**, **Need** and **Situation** in positional order,
2. templates of type **Topic of Document** sorted in descending order of **Weight**,
3. templates of type **Possible Topic** sorted in descending order of **Weight**,
4. templates of type **Topic Description**, **Study**, **Interest**, **Development**, **Entity Introduction**, **Research Goal**, **Conceptual Goal**, **Conceptual Focus** and **Focus** in positional order,
5. templates of type **Method** and **Experiment** in positional order,
6. templates of type **Results**, **Inference**, **Knowledge** and **Summarization** in positional order,
7. templates of type **Entity Identification** in positional order,
8. templates of type **Topic of Section** in section order, and
9. templates of type **Signaling Structural** and **Signaling Concepts** in positional order.

This step grouped information of the same type together and this is done in order to achieve coherence at the conceptual level. The result is the list **sorted content**.

In order to implement the **Merge** described in Section 4.4 we do the following: sequences of templates of type **Topic of Document**, **Topic of Section**, and **Signaling Structural or Concepts** on the **sorted content** are replaced by a **Merged** template, which is a combination of single and multi templates. This process dynamically couples in a **Multi** template those templates containing the same citation form in the slot **Predicate** and assembles multi and single templates of the same type in a merged template allowing the presentation of topical information in complex sentences instead of single ones (e.g., “**Presents X and Y.**” instead of “**Presents X. Presents Y.**” and “**Presents X; describes Y; and also discusses Z.**” instead of “**Presents X. Describes Y. Discusses Z.**”)

\textsuperscript{2}Note, that interesting information about a potential topic can be overlooked in this process (i.e., domain specific information, productive patterns not present in the model, etc.).
5.7. GENERATION

<table>
<thead>
<tr>
<th>Section 2 describes HuDL in greater detail and section 3 discusses system integration and the IMA.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type:</strong></td>
</tr>
<tr>
<td><strong>Id:</strong></td>
</tr>
<tr>
<td><strong>Predicate:</strong></td>
</tr>
<tr>
<td><strong>Section:</strong></td>
</tr>
<tr>
<td><strong>Argument:</strong></td>
</tr>
<tr>
<td><strong>Position:</strong></td>
</tr>
<tr>
<td><strong>Topic candidates:</strong></td>
</tr>
<tr>
<td><strong>Weight:</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section 2 describes HuDL in greater detail and section 3 discusses system integration and the IMA.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type:</strong></td>
</tr>
<tr>
<td><strong>Id:</strong></td>
</tr>
<tr>
<td><strong>Predicate:</strong></td>
</tr>
<tr>
<td><strong>Section:</strong></td>
</tr>
<tr>
<td><strong>Argument:</strong></td>
</tr>
<tr>
<td><strong>Position:</strong></td>
</tr>
<tr>
<td><strong>Topic candidates:</strong></td>
</tr>
<tr>
<td><strong>Weight:</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>An example implementation is given in section 4 and section 5 contains the conclusions.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type:</strong></td>
</tr>
<tr>
<td><strong>Id:</strong></td>
</tr>
<tr>
<td><strong>Predicate:</strong></td>
</tr>
<tr>
<td><strong>Section:</strong></td>
</tr>
<tr>
<td><strong>Argument:</strong></td>
</tr>
<tr>
<td><strong>Position:</strong></td>
</tr>
<tr>
<td><strong>Topic candidates:</strong></td>
</tr>
<tr>
<td><strong>Weight:</strong></td>
</tr>
</tbody>
</table>

The three templates were already selected and sorted and so, they are merged together for the process of regeneration

| **Type:** | merged |
| **Templates:** | 18, 17, 19 |

**Fig. 5.32:** Instantiation of the Topic of Section Templates and Merging

In Figure 5.32, we exemplify the process of merging three templates of type **Topic of Section** that were instantiated with different sentences of the source document (before each template we give the sentence used for instantiation). As the three templates contain different verbs (e.g., “describe”, “discuss” and “give”) no multi template is produced in this case. For an example in which three templates are grouped together in a multi template refer to Section 5.9.

The sorted templates and the merged templates constitute the **text plan.** This kind of strategy for text summarization could produce a text presenting the information in a different order than the one in which the information appeared in the source text. This is particularly important in summarization of technical texts: the topical information could perfectly appear in the conclusion of the paper while it has to be almost always the first information to report in the abstract after the background information. To our knowledge our approach is unique in this particular aspect.
**Presentation**: The text plan will produce a text paragraph. Each element (template or merged template) in the text plan is used to produce a sentence. The sentences will be concatenated and the connective between them is the full stop. The schema of presentation of a text plan composed of $n \geq 1$ templates $Tmpl_i$ is as follows:

$$Text = \Theta_{i=1}^{n}[Tmpl_i \oplus "]$$

The notation $\Theta$ means the string produced by the generation of $A$ and the expression $\Theta_{i=1}^{n} A_i$ stands for the concatenation of $A_i$. We assume that all the parameters necessary for the generation are available (i.e., voice, tense, number, position, etc.).

The structure of the sentence depends on the type of template holding it. Information of type **Situation, Problem Solution, Need for Research**, etc. is reported as in the original document with few modifications (concept re-expression). For example, the generation of a sentence from the **Problem Identification** template consists of the presentation of the information in the **Content** slot of that template. But other types may require additional reformulation in order to implement the transformations observed in Section 4.4.

For the **Topic of Document** template, the generation procedure is as follows: the verb form for the predicate in the **Predicate** slot is generated in the present tense, 3rd person of singular in active voice at the beginning of the sentence because we are following the patterns observed in Chapter 4; and the parsed sentence fragment from the **What** slot is generated in the middle of the sentence. Some elements in the parsed sentence fragment require reformulation while others are presented as they were found in the source document. The latter includes prepositions, cue phrases, verb groups and some noun groups without correlation with the conceptual model. Instead, the concepts **author, research paper and author related** are presented using a pre-defined expression (e.g., “the authors”, “the article”). Noun groups introduced by demonstrative are also reformulated as shown in Section 4.4 (e.g., “This X”, “That X” are presented as “The X”). Elements at the beginning of the sentence are capitalized and an argument presented in the middle of the sentence is transformed into lower case. Exceptions are proper nouns that are presented always as they were found in the text. If **SUM** detects an acronym without expansion it also presents its expansion and records that fact to avoid repetitions. As the templates contain parsed sentence fragments, the correct punctuation has to be generated. The schema of presentation of a template $Tmpl$ of type **Topic of Document** is:

$$Tmpl = Tmpl.Predicate \oplus Tmpl.What$$

This kind of schema avoids the generation of sentences like “X will be presented”, “X have been presented” or “We have presented here X” which are usually found on source documents but which are awkward in the context of the abstract text-type. This standardization of the verbs implements the **Syntactic Verb Transformation** reported in Section 4.4 (see Figure 4.12).
5.7. GENERATION

For the **Possible Topic** template, if the Continuation slot is empty the generation procedure is the same as for the topic. Otherwise, the procedure is as follows: the parsed sentence fragment from the Argument slot is generated at the beginning of the sentence; the verb form for the predicate in the Predicate slot is generated in the present tense in the passive voice (using the information about the grammatical number from the verb (is presented vs. are presented)) in the middle of the sentence; and the parsed fragment from the Continuation slot is generated in the middle of the sentence. The schema of presentation of a template $Tmpl$ of type **Possible Topic** in this case is:

$$Tmpl = Tmpl.\text{Argument} \oplus Tmpl.\text{Predicate} \oplus Tmpl.\text{Continuation}$$

The schema of generation of a **Multi** template is as follows: the verb form for the predicate in the Predicate slot is generated in the present tense, 3rd person of singular in active voice at the beginning of the sentence; and the argument is presented as a conjunction. The schema of sentence production of a template $Tmpl$ of type **Multi** is:

$$Tmpl = Tmpl.\text{Predicate} \oplus Tmpl.\text{Arguments}$$

The schema of presentation of the slot Arguments of a **Multi** template is as follows:

1. If $Tmpl.\text{Arguments}$ contains two elements the schema is:

$$Tmpl.\text{Arguments} = Tmpl.\text{Arguments}_1 \oplus \text{“and”} \oplus Tmpl.\text{Arguments}_2$$

2. If $Tmpl.\text{Arguments}$ contains $n(\geq 3)$ elements the schema is:

$$Tmpl.\text{Arguments} = (\oplus_{i=1}^{n-1}[Tmpl.\text{Arguments}_i \oplus \text{“;”}]) \oplus \text{“and”} \oplus Tmpl.\text{Arguments}_n$$

The schema of generation of a **Merged** template $Tmpl$ is:

$$Tmpl = (\oplus_{i=1}^{n-1}[Tmpl.\text{Templates}_i \oplus \text{“;”}]) \oplus \text{“and also”} \oplus Tmpl.\text{Templates}_n$$

For the example shown in Figure 5.32, the output will be “Describes HuDL in greater detail; discusses system integration and IMA; and also gives an example implementation.”

This dynamic combination of information in order to produce a better text is another new aspect of our approach to text summarization. In the actual implementation we merge up to three templates together. The schemas of sentence reformulation are presented along with examples of input and regenerated sentences in Tables 5.3-5.8. Note that each template dictates the form to be used for the reformulation of the verb on the Predicate slot.

The elaboration of the topics is presented upon reader’s demand (informative generation). **SumUM** selects the informative templates from the informative database linked to the topics the reader have chosen, sorts the information using the Position slot and generates the sentences using the schema **Informative** (Table 5.8).
<table>
<thead>
<tr>
<th><strong>Situation Schema</strong></th>
<th><strong>Template Situation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Robotics approaches have been applied since the late 1970s with more and more advanced devices and strategies.</td>
<td>Robotics approaches have been applied since the late 1970s with more and more advanced devices and strategies.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Problem/Solution Schema</strong></th>
<th><strong>Template Content</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Taking into account the extreme complexity of the problems related to the environment (light conditions, sizes and shapes, overlapping fruits, branches and leaves, colors) and the limitations of the current approaches, the implementation of a fully automatic and real-time solution for this task seems far away.</td>
<td>Taking into account the extreme complexity of the problems related to the environment and the limitations of the current approaches, the implementation of a fully automatic and real-time solution for this task seems far away.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Need Schema</strong></th>
<th><strong>Template Content</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Some remaining modifications to the detaching tool are required to reduce the detaching cycle time which is currently about 2s.</td>
<td>Some remaining modifications to the detaching tool are required to reduce the detaching cycle time which is currently about 2s.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Entity Introduction Schema</strong></th>
<th><strong>Template Entity + Template What</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Industrial background has been traditionally the application field for robotics.</td>
<td>Industrial background has been traditionally the application field for robotics.</td>
</tr>
</tbody>
</table>

Tab. 5.3: Schemas of Sentence Reformulation for templates of type Situation, Problem/Solution, Need, and Entity Introduction
<table>
<thead>
<tr>
<th>Topic of Document Schema</th>
<th>( Tmpl.Predicate @ Tmpl.What ). We use the simple present for the verb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>We report the implementation of a high-level interpretation module that is able to recognize complex actions from low-level physical events in the virtual world, and we discuss its performance as well as directions for further developments.</td>
<td>Reports the implementation of a high-level interpretation module that is able to recognize complex actions from low-level physical events in the virtual world.</td>
</tr>
<tr>
<td>Topic Description Schema</td>
<td>( Tmpl.Description )</td>
</tr>
<tr>
<td>The technology described in this paper was developed during Amadeus phase 1 and a single gripper system demonstrated in a laboratory environment.</td>
<td>The technology was developed during Amadeus phase 1 and a single gripper system demonstrated in a laboratory environment.</td>
</tr>
<tr>
<td>Possible Topic Schema</td>
<td>( Tmpl.Predicate \oplus Tmpl.Argument ). We use the simple present for the verb.</td>
</tr>
<tr>
<td>A method for explicitly manipulating contextual information during deduction is proposed, where pronouns are resolved against this context during deduction.</td>
<td>Proposes a method for explicitly manipulating contextual information during deduction.</td>
</tr>
<tr>
<td>Author Development Schema</td>
<td>( Tmpl.Who \oplus Tmpl.Predicate \oplus Tmpl.Argument ). We use the simple present or simple past for the verb.</td>
</tr>
<tr>
<td>We have developed a regional ontology of the domain of electrical network planning in order to use it within a technical documentation consultation system.</td>
<td>The authors developed a regional ontology of the domain of electrical network planning in order to use it within a technical documentation consultation system.</td>
</tr>
</tbody>
</table>

Tab. 5.4: Schemas of Sentence Reformulation for templates of type Topic of Document, Topic Description, Possible Topic, and Author Development
<table>
<thead>
<tr>
<th>Development Schema</th>
<th>$\text{Template.Argument} \oplus \text{Template.Predicate} \oplus \text{Template.Continuation}$. We use the same tense as in the source document.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The now-patented solutions were developed at MIT in support of the WAM project.</td>
<td><em>The now-patented solutions were developed at MIT in support of the WAM (the whole-arm manipulator) project.</em></td>
</tr>
<tr>
<td>Author Study Schema</td>
<td>$\text{Template.Who} \oplus \text{Template.Predicate} \oplus \text{Template.Argument}$. We use simple present or simple past when the subject is the author. We use the simple present when the subject is the research paper. Otherwise, we use the tense of the source document.</td>
</tr>
<tr>
<td>Since 1994, we have investigated the behaviour modelling of electronic ANNs with global perturbation conditions.</td>
<td><em>The authors investigated the behavior modelling of electronic ANNs with global perturbation conditions.</em></td>
</tr>
<tr>
<td>Study Schema</td>
<td>$\text{Template.Argument} \oplus \text{Template.Predicate} \oplus \text{Template.Continuation}$. We use the original tense.</td>
</tr>
<tr>
<td>Here the hybrid approach is evaluated indirectly, on the task of tag prediction.</td>
<td><em>The hybrid approach is evaluated.</em></td>
</tr>
<tr>
<td>Author Interest Schema</td>
<td>$\text{Template.Who} \oplus \text{Template.Predicate} \oplus \text{Template.Argument}$. We use the original tense.</td>
</tr>
<tr>
<td>In the work presented, we are concerned with the extraction of meta-knowledge from the Web.</td>
<td><em>The authors are concerned with the extraction of meta-knowledge from the Web.</em></td>
</tr>
</tbody>
</table>

Tab. 5.5: Schemas of Sentence Reformulation for templates of type Development, Author Study, Study, and Author Interest
### Conceptual Goal Schema

The objective of the Aristotle project is to build an automatic data system that is capable of producing a semantic representation of the text in a canonical form.

### Research Goal Schema

Thus, the HRP project aims to develop a safe and reliable human friendly robot system capable of carrying out complicated tasks and supporting humans both at work and in leisure activities.

### Conceptual Focus Schema

The focus of this paper is to consider usability issues in catalogs independently of current trends on the Internet.

### Focus Schema (for author and research paper)

We focus on the problem of pronoun resolution and the way in which it complicates automated theorem proving for natural language processing.

### Focus Schema (for other concepts)

In the last few years, several researches have focused on the possibility for distributed systems to host mobile and dynamic entities.

<table>
<thead>
<tr>
<th>Template Type</th>
<th>Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual Goal</td>
<td>$\text{tmpl.marker} \oplus \text{tmpl.predicate} \oplus \text{tmpl.goal}$</td>
</tr>
<tr>
<td>Research Goal</td>
<td>$\text{tmpl.entity} \oplus \text{tmpl.predicate} \oplus \text{tmpl.goal}$</td>
</tr>
<tr>
<td>Conceptual Focus</td>
<td>$\text{tmpl.marker} \oplus \text{tmpl.predicate} \oplus \text{tmpl.focus}$</td>
</tr>
<tr>
<td>Focus</td>
<td>$\text{tmpl.predicate} @ \text{tmpl.focus}$</td>
</tr>
<tr>
<td>Focus (for other concepts)</td>
<td>$\text{tmpl.marker} \oplus \text{tmpl.predicate} \oplus \text{tmpl.focus}$</td>
</tr>
</tbody>
</table>

Tab. 5.6: Schemas of Sentence Reformulation for templates of type Conceptual Goal, Research Goal, Conceptual Focus, and Focus
<table>
<thead>
<tr>
<th>Method Schema</th>
<th>Template.Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storyboarding is a technique designed to generate consensus and closure via a tangible, interactive system concept.</td>
<td>Storyboarding is a technique designed to generate consensus and closure via a tangible, interactive system concept.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment Schema</th>
<th>Template.Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>In these experiments the robot learned to recognise landmarks (boxes, light, aluminium covered ground areas).</td>
<td>In the experiments the robot learned to recognise landmarks.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Result Schema</th>
<th>Template.Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>The results of this experience show that it is possible to realize a cooperative manipulation system with standard industrial robots, not prototypes, without re-designing the controller hardware.</td>
<td>The results of this experience show that it is possible to realize a cooperative manipulation system with standard industrial robots, not prototypes, without re-designing the controller hardware.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inference Schema</th>
<th>Template.Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our study showed that extending the four most important analysis and transformation techniques traditionally used for vectorization leads to significant increases in speedup.</td>
<td>The study showed that extending the four most important analysis and transformation techniques traditionally used for vectorization leads to significant increases in speedup.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author Knowledge Schema</th>
<th>Template.Who ⊕ Template.Predicate ⊕ Template.What. We use the tense found on the source document.</th>
</tr>
</thead>
<tbody>
<tr>
<td>This paper explains the internal design of the Preci-Check measuring system and will discuss the collaboration of hardware and software.</td>
<td>The paper explains the internal design of the Preci-Check measuring system.</td>
</tr>
</tbody>
</table>

Tab. 5.7: Schemas of Sentence Reformulation for templates of type Method, Experiment, Result, Inference, and Author Knowledge
### Summarization Schema

The approach presented combines static information with actual execution information to produce views that summarize the relevant computation.

### Entity Identification Schema

This database contains optimised parameters obtained by studies and extensive experimentation.

### Topic of Section Schema

The details of the GUI are discussed with respect to specific examples in section 5.0.

### Signaling Structural Schema

Table I Important welding process parameters shows the fields in the welding database.

### Signaling Concept Schema

Walk-through teaching method

### Informative Schema

The Agribot has been developed at Instituto de Automtica Industrial to cope with the outlined requirements.

<table>
<thead>
<tr>
<th><strong>Summarization Schema</strong></th>
<th><strong>tmpl.Content</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>The approach presented combines static information with actual execution information to produce views that summarize the relevant computation.</td>
<td><strong>tmpl.Content</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Entity Identification Schema</strong></th>
<th><strong>tmpl.Entity ⊕ tmpl.What</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>This database contains optimised parameters obtained by studies and extensive experimentation.</td>
<td><strong>tmpl.Entity ⊕ tmpl.What</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Topic of Section Schema</strong></th>
<th><strong>tmpl.Predicate ⊕ tmpl.Argument.</strong> We use the simple present.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The details of the GUI are discussed with respect to specific examples in section 5.0.</td>
<td><strong>tmpl.Predicate ⊕ tmpl.Argument.</strong> We use the simple present.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Signaling Structural Schema</strong></th>
<th><strong>tmpl.Predicate ⊕ tmpl.Argument.</strong> We use the simple present.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table I Important welding process parameters shows the fields in the welding database.</td>
<td><strong>tmpl.Predicate ⊕ tmpl.Argument.</strong> We use the simple present.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Signaling Concept Schema</strong></th>
<th><strong>tmpl.Predicate ⊕ tmpl.What</strong>. We use the simple present.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk-through teaching method</td>
<td><strong>tmpl.Predicate ⊕ tmpl.What</strong>. We use the simple present.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Informative Schema</strong></th>
<th><strong>tmpl.Content.</strong> The information is stated as in the source document without transformation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Agribot has been developed at Instituto de Automtica Industrial to cope with the outlined requirements.</td>
<td><strong>tmpl.Content.</strong> The information is stated as in the source document without transformation.</td>
</tr>
</tbody>
</table>

**Tab. 5.8:** Schemas of Sentence Reformulation for templates of type **Summarization, Entity Identification, Topic of Section, Signaling Structural, Signaling Concept, and Informative**
5.8 A short annotated example

In Figure 5.33, we show an example of the type of abstract that SumUM produces. The source document is the text presented in Figure 5.2. It is a document of 13K characters and contains 97 sentences and 1868 words. It is organized in four sections: “Introduction”, “REVERSA”, “ModelMaker” and “Conclusions”. The indicative abstract (which is less than 1% of the original text) was produced by reformulating the sentence (1) which is the first sentence of the last section.

(1) Two non-contact scanning systems, REVERSA and ModelMaker have been described and their application in industry demonstrated.

Abstract Introducing the Topics

Describes two non-contact scanning systems, REVERSA and ModelMaker.

Topics: CADCAM system - ModelMaker - REVERSA - standard dual view system - system

Fig. 5.33: Indicative abstract and list of topics produced by SumUM for the source document “Features 3D scanning systems for rapid prototyping” from the Journal Assembly Automation, 17(3), 1997

The indicative abstract shown in Figure 5.33 contains a list of topics. Information about them is presented upon user demand. Figure 5.34, presents all the “elaborations” the system found about REVERSA and ModelMaker these include: definition, advantage and use³.

³Note that in the source document there is more information about the topics that SumUM was not able to identify due to limitations in its implementation.
Information about the Topics REVERSA and ModelMaker

The first sensor, REVERSA, is a standard dual view system which can be mounted on a purpose built four-axis translation stage or retro-fitted to third party CNC and CMM (the co-ordinate measuring machine) machines.

REVERSA is a dual viewpoint non-contact laser scanner which comes complete with scanning software and data manipulation tools.

The ModelMaker scanning system is a combination of a 3D laser stripe sensor, 6 DoF position localizer and a PC in which up to two interface cards are mounted.

ModelMaker can simply be retrofitted to existing arms providing the benefits of a portable CMM with dense depth data sets.

Use of ModelMaker allowed the part to be scanned in under one hour, a job that might have taken days using the traditional point to point touch process, and did deform the model in the process.

Fig. 5.34: Topic elaboration for the abstract in Figure 5.33

This abstract was produced as follows:

- During the pre-processing and interpretation step SumUM identifies the four sections of the article and the titles. All the sentences are syntactically and semantically analyzed, in particular in sentence (1), SumUM identifies the verb groups scanning, have been described and demonstrated and the noun groups two non-contact\(^4\), systems, REVERSA, ModelMaker, their application and industry. The verb group have been described is interpreted as a make known relation and demonstrated as an infer relation. The noun group their application is interpreted as an instance of the application concept.

Sentence (1) is classified as: (i) Possible Topic because of its position in the last section of the document and because of the make known relation, (ii) Inference because of the infer relation, and as (iii) Usefulness of Entity because of the application concept. All this information is recorded in the conceptual index. The topical structure is composed of the following elements: feature, scan system, rapid prototyping, introduction, REVERSA, ModelMaker, and conclusion.

- During the indicative selection different templates are instantiated. Sentence (1) instantiates a template of type Possible Topic which is shown in Figure 5.35. This is done by extracting the syntactic information about the make known relation and its argument which is the string to the left of the relation. The string to its right is ignored because of the presence of the conjunction (and) next to it. Note that the fillers of the slots Argument and Continuation are in fact parsed sentence fragment that here are

\(^4\)The tagger interpreted the word non-contact as a noun.
SumUM matches each term of the topical structure with the templates in the IDB. The template Possible Topic matches the terms scan system, REVERSA and ModelMaker. The terms feature, rapid prototyping, introduction and conclusion do not have matched templates. As the content of the Weight slot of the Possible Topic template is the greatest among all those matching the topical structure, only this template is incorporated in the indicative content. The potential topics are non-contact, system, REVERSA and ModelMaker (obtained from the Topic candidates slot) and their expansions CADCAM system, arm system, scan system, second system, standard dual view system and table system, which were found on the term tree.

- During the informative selection SumUM considers “informative” sentences containing the potential topics. For example, the potential topic REVERSA appears in the following sentence:

(2) REVERSA is a dual viewpoint non-contact laser scanner which comes complete with scanning software and data manipulation tools.

The type of this sentence is Definition because it contains the define relation. In addition, the sentence matches a pattern of definition which asks for the potential topic to appear immediately to the left of the define relation and a noun group to its right. This sentence instantiates a template of type Definition shown in Figure 5.36, only structural elements (figures, etc.) are removed from the sentence. The potential topic REVERSA became a topic of the document and the Definition template is included in the informative data base.

- In order to generate the indicative abstract, SumUM uses the schema available to present the text plan. In this case, the text plan contains a single template Possible Topic: SumUM produces a sentence from that template and then a full stop. In order to produce the sentence it uses the schema for the Possible Topic template (Section 5.7). The Continuation is empty, thus, the verb describe is expressed in the present tense, 3rd person of singular in active voice at the beginning of the sentence (Describes). The argument (Argument slot) is generated in the middle of the sentence (two non-contact systems, REVERSA and ModelMaker). This step gives the abstract shown in Figure 5.33.

- In order to obtain the informative part of the abstract for the topics REVERSA and ModelMaker, SumUM retrieves from the informative data base templates containing those terms as topics. The templates are sorted according to its position and the information in the Content slot is presented.
5.9. A LONGER EXAMPLE

<table>
<thead>
<tr>
<th>Type</th>
<th>possible topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>1</td>
</tr>
<tr>
<td>Predicate</td>
<td>describe, ...</td>
</tr>
<tr>
<td>Argument</td>
<td>Two non-contact scanning systems, REVERSA and ModelMaker</td>
</tr>
<tr>
<td>Continuation</td>
<td>none</td>
</tr>
<tr>
<td>Position</td>
<td>Sentence 1 from Section 4</td>
</tr>
<tr>
<td>Topic candidates</td>
<td>non-contact, system, REVERSA, ModelMaker</td>
</tr>
<tr>
<td>Weight</td>
<td>43</td>
</tr>
</tbody>
</table>

FIG. 5.35: Instantiated Possible Topic Template

<table>
<thead>
<tr>
<th>Type</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>41</td>
</tr>
<tr>
<td>Topic</td>
<td>REVERSA</td>
</tr>
<tr>
<td>Predicate</td>
<td>be, ...</td>
</tr>
<tr>
<td>Content</td>
<td>REVERSA is a dual viewpoint non-contact laser scanner which ...</td>
</tr>
<tr>
<td>Position</td>
<td>Sentence 1 from Section 2</td>
</tr>
</tbody>
</table>

FIG. 5.36: Instantiated Definition Template

Note that the indicative abstract introduces the two main entities the source document talks about, but very little is known about those entities from the indicative abstract. The informative abstract fulfills that purpose. The drawback is that other entities are as well elaborated in the source document, but as they are not present neither in the titles nor in the sentences selected by SumUM, they are not taken in consideration.

5.9 A longer example

In Figures 5.37 and 5.38, we present another example of automatic abstract. The source document (partially shown in Figure 5.39) is 31K characters and contains 4166 words.

The indicative abstract represents about 4% of the source document. The information that SumUM includes in the abstract and its order is: the **Topic of Document** (three templates are chosen and merged in a Merged template), the **Author Development** (two templates are selected), the **Entity Introduction** (one template is selected), and the **Research Goal** (one template is selected).

The text plan is composed of four templates: one Merged template obtained by merging the three Topic templates sharing the same predicate, two Author Development templates, one Entity Introduction template and one Research Goal template. SumUM presents the topical information following the pattern dictated by the Merged tem-
Abstract Introducing the Topics

Presents a novel connectionist architecture, DRAMA (a dynamical recurrent associative memory); a new implementation of DRAMA in an autonomous doll-shaped robot, which is taught a synthetic proto-language; and experiments, in which the robot’s imitative skills are used for teaching the robot a language. The authors implemented the model in a number of experiments with wheeled robots, where the robot learned spatial regularities of the environment by recognising landmarks and temporal regularities by recording the time delay for travelling from one landmark to the other. The authors implemented the DRAMA architecture in a number of experiments with wheeled LEGO or FISCHERTECHNIK robots which are widely used tools for research on mobile robots. Online learning is a fundamental requirement to achieve robust and adaptable robots. The authors aimed for a general robot control mechanism, as independent as possible from the particular environment and hardware used for the implementation, ANNs (artificial neural networks) were more interesting to the authors than RL (reinforcement learning) and EA (evolutionary algorithms) techniques.

Topics: control architecture - doll robot - DRAMA - DRAMA architecture - DRAMA model - implementation - language - learning - model - robot - teacher robot

Fig. 5.37: Indicative abstract and list of topics produced by SumUM for the source document “DRAMA, a connectionist architecture for online learning and control of autonomous robots: experiments on learning of a synthetic proto-language with a doll robot” from the Journal Industrial Robots 26(1), 1999.

It expands the acronym DRAMA the first time it appears. Each template of type Development is presented in an independent sentence following the pattern of Author Development. The information from the template Entity Introduction is presented as in the source document. When SumUM presents the information associated to the Research Goal template, it expands the acronyms ANNs, RL and EA. The concept author is reformulated with the expression “the authors” (three times, this could be improved introducing pronominal references for example or by trying a merge between informations of templates of type Author Development as well).

This abstract is more verbose than the previous one, but it has some drawbacks. The expressions:

...for teaching the robot a language...

and

...robot, which is taught a synthetic proto-language...

indeed refer to the same conceptual information. The same is true for the expressions:

...
Information about the topics

**DRAMA** is a fully connected network with self-connections on each unit.

The **DRAMA architecture** shows characteristics which make it very relevant for all types of autonomous robotic agents.

The **robots** are provided with a micro-controller with 512k byte EPROM space and 128k byte Static RAM.

In the paper, a novel implementation of the **DRAMA model** into a doll-shaped robot was presented, for teaching the robot a synthetic proto-language.

Imitation, communication and **learning** are important skills to possess by a robot expected to interact with humans through daily tasks.

The **doll robot** is provided with the three competencies, which are completely controlled by the DRAMA architecture.

---

**FIG. 5.38:** Informative abstract produced by **SumUM** for the source document “DRAMA, a connectionist architecture for online learning and control of autonomous robots: experiments on learning of a synthetic proto-language with a doll robot” from the Journal Industrial Robots 26(1), 1999

...experiments with wheeled LEGO or **FISCHERTECHNIK robots**

and

experiments with wheeled robots.

But to deduce that those expressions are in some sense “equivalent” or elaborations of each another, we would need world knowledge. We did not address this issue of the repetition of information when it is not stated verbatim.
[...] 

Introduction

There is an high potential in terms of both industrial applications [...] 

This paper presents experiments, in which the robot’s imitative skills are used for teaching the robot a language (information used for the first sentence of the abstract) [...] 

Because we aimed for a general robot control mechanism, as independent as possible from the particular environment and hardware used for the implementation, ANNs were more interesting to us than RL and EA techniques. (information used for the last sentence of the abstract) [...] 

Online learning is a fundamental requirement to achieve robust and adaptable robots. (information used for the forth sentence of the abstract) [...] 

We implemented the model in a number of experiments with wheeled robots, where the robot learned spatial regularities of its environment by recognising landmarks and temporal regularities by recording the time delay for travelling from one landmark to the other. (information used for the second sentence of the abstract) [...] 

This paper presents a new implementation of DRAMA in an autonomous doll-shaped robot, which is taught a synthetic proto-language (information used for the first sentence) [...] 

The robot’s controller 

The DRAMA network 

The associative module of the robot’s controller is made of an artificial neural network, where each unit is [...] 

Experiments on landmarks recognition and labelling

We implemented the DRAMA architecture in a number of experiments with wheeled LEGO or FISCHER-TECHNIK robots which are (information used for the third sentence) [...] 

Experiments on teaching a doll robot 

The DRAMA architecture was implemented in a doll-shaped robot [...] 

Conclusion

This paper presented a novel connectionist architecture, DRAMA (information used for the first sentence) [...] 

References

Billard A. (1998), "DRAMA, a connectionist model for robot learning : experiments on grounding [...] 

Fig. 5.39: Source Document for the Abstract in Figure 5.37
5.10 Discussion

We worked with the hypothesis that “Introduction” and “Conclusion” are conceptual sections, thus specific types of conceptual information are to be reported there such as the statement of the situation, the problem, the solution, the topic of the document, the structure of the article, etc. This is a reasonable assumption arising from studies in genre analysis (Bhatia, 1993; Jordan, 1993; Paice, 1991) which studied rhetoric-conceptual moves in formal research articles. We also assume that titles are indicative of the content, this assumption is acceptable in the context of a technical article where authors will use a number of strategies in order to explicitly mark the topics including the use of titles Jordan (1996). The assumption that the title indicates the content has also been used in work in automatic indexing in particular in order to create indexes (Borko and Bernier, 1978 and in text summarization (see Section 4.2).

We are assuming that the articles do not necessarily follow a prototypical structure, this is because we think that articles with a prototypical structure such as “Introduction”, “Method”, “Results”, “Analysis”, “Previous Work”, “Discussion”, “Conclusion” promote specific summarization strategies. For example, specific information extraction algorithms could be used in order to extract the “main” results from the “Results” section, to obtain the complete description of the “new” procedures and the description of the experimental subjects from the “Method” section, or even the “relevant” relation between the work of the author and previous works from the “Previous Work” section. That information could be integrated in a new text. However, for this kind of article the main title could eventually be used in order to select information for the statement of the situation and for the topical sentence.

In the process of reading and comprehension, humans are able to deduce the topic of a discourse by generalization of a few examples among other cognitive processes. For example, a reader could generalize the concepts “cat” and “dog” in the appropriate context into the concept “pet” or from “Michelangelo” and “Leonardo” deduce the concept “Renaissance.” But, in order to achieve this intelligent behavior, natural language processing systems require extensive knowledge. We have chosen to represent the topics of a document using the terms explicitly stated in the source document and so, our approach is superficial and only an approximation to the real concept of topic considered in studies in Cognitive Psychology (Kieras, 1982) and others.

In our approach, the selection of templates for the indicative abstract is based on a combination of relevance of the information and a matching process. Two aspects of relevance were considered here: first, the status of the information motivated by the analysis of the corpus; second, the relevance of the terms approached by the classical method of term distribution on a single document. Other approaches have already used this kind of strategy. For example, Paice (1981) and Paice and Jones (1993) concentrated on indicative information on technical articles but they did not try to further extract the information from the sentences in order to construct a new expanded abstract. However, they have addressed the issue of expression in text summarization and the relevance of the extracted concepts based
on some heuristic criteria (concept repetition). Lehmann (1997), based his approach on the selection of specific types of information in the context of the technical article. The relevance of the information is based on an a priori score (empirically obtained) associated with lexical expressions of this text-type and not on the content of the text.

In future work we will explore additional processes of content selection, in particular we will study how information elaborated in the source document can be selected even if overlooked by the indicative selection. This is motivated by some observation on the corpus: abstractors produce sentences like Describes Topic when in the source document they found the actual description of Topic. The reader understands that a text segment is a description of Topic and reformulates that information with a “macro proposition.” This motivates the study of automatic text classification techniques in order to categorize text fragments which is the first step towards the production of such intelligent behavior.

Our matching process operates between terms from titles and terms from templates and is based on the substring relation. This process could probably be improved considering that two terms match if a lexical relation like synonymy exists between them; this could be checked using a lexical data base like WordNet (Fellbaum, 1998). But in order to verify that hypothesis, additional implementation and experimentation is needed.

In the matching process, the indicative selection only considers the indicative templates and the titles to calculate the content of the indicative abstract. However, we believe that a better result could be obtained by considering a matching process between the indicative templates and the informative templates obtained independently of the potential topics as an alternative to the exclusive use of terms from titles. This hypothesis also requires additional experimentation in order to be validated.

SumUM only analyses sentences and does not consider relations that cross sentence boundaries. This is not a limitation of the model, instead it is a pragmatic decision we made when designing the architecture of SumUM. Topic elaboration is only based on term repetition, a common phenomenon in long technical documents (Justeson and Katz, 1995). Other cohesive and coherence phenomena like anaphora and synonymy were not addressed in the present work and will be subject of future improvements.

Here, we addressed the expansion of the indicative abstract by considering that the intended and unknown reader would be interested in general elaborations about the topics and we have assumed that there are patterns of language production that can be used to identify those types of information. Patterns are not the ultimate solution for text classification. They are ambiguous and unreliable because they consider sentences out of context. In future work we will explore more robust techniques of text classification.

We have addressed the issue of reformulation, usually neglected in domain independent summarization. Our approach to text (re)generation is based on the arrangement of the templates using a conceptual order and on the specification of a few schemas of text produc-
tion. The schemas of text production implement some aspects of the Merge transformation studied in Chapter 4 (we have only considered the merge of “topical” information). Radev and McKeown (1998), have addressed the issue of merging of information in the context of multi-document summarization. The merging is achieved by the implementation of summary operators that integrate the information of different templates from different documents referring to the same event. While those operators are dependent on the specific task of multi-document summarization and to some extent of the particular domain they deal with, it is interesting to observe that some of their ideas could be applied in order to improve our texts. For example, their “refinement” operator could be used to improve the descriptions of the entities of the indicative abstract shown in Figure 5.33. The entities from the indicative abstract could be refined with definitions or descriptions from the informative data base in order to obtain a better and more compact text such as:

_Describes two non-contact scanning systems: REVERSA, which is a dual viewpoint non-contact laser scanner, and the ModelMaker scanning system, which is a combination of a 3D laser stripe sensor, 6 DoF position localizer and a PC._

This kind of transformation has recently been addressed by Mani et al. (1999) using syntactic analysis. While full syntactic analysis is not present in our actual model, some kind of partial parsing will be explored in future work.

Some aspects of editing of the source material presented in the previous chapter are present in Selective Analysis. Syntactic verb transformation is incorporated on some schemas of presentation; verb selection is achieved by using domain verbs (show, overview, etc.) to introduce domain concepts and descriptions found on structural elements; concept re-expression is achieved using fixed lexical forms to express domain concepts; acronym expansion was also considered while most of the abstract is presented with the “words of the author.” The split and delete transformations are achieved by the process of information extraction that considers only parts of sentences in order to instantiate templates. Marcu (1997a) has already addressed the “safe” deletion of clause components using Rhetorical Structure Theory. His approach relies on discourse markers in order to segment sentences and construct text structures. It would be interesting to incorporate some aspects of his model to improve our process of extraction and deletion. For example, rhetorical parsing could be applied to sentences in order to produce less verbose expressions. For example, with rhetorical parsing we could produce the sentence:

_The authors implemented the model in a number of experiments with wheeled robots._

from the third sentence of the abstract in Figure 5.37 by eliminating the clause introduced by where.

While this discussion addresses theoretical and practical limitations of this work, the abstracts produced by SumUM have been already evaluated by human informants. The evaluations are described in Chapter 7.
5.11 Summary

In this chapter we have presented Selective Analysis, a method for the summarization of scientific and technical texts. The idea in our approach is that the topics of a technical document can be identified using a limited set of specific concepts and relations; these can eventually be expanded using domain independent relations such as definition and description.

Our method was deeply influenced by the results of our corpus study, nevertheless, it has many points in common with recent theoretical and programmatic directions in automatic text summarization. On one hand, Spark Jones (1997, 1999) argues in favor of a kind of “indicative, skeletal summary” and the need to explore dynamic, context-sensitive summarization in interactive situations where the summary changes according to the user needs. On the other hand, Hutchins (1995) advocates for indicative summaries, produced from parts of the document where the topics are likely to be stated. These abstracts are well suited for situations in which the actual user is unknown (i.e., a general reader), and so the abstract will provide the reader with good entry points for the information. If the users were known, the abstract could be tailored towards their specific profiles (profiles could include the specification of readers interested in types of information like conclusions, definitions, methods, or user needs expressed in a “query” to an information retrieval system (Tombros et al., 1998)), but our method was designed without any particular reader in mind and with the assumption that a text does have a “main” topic. Our method also mirrors somehow professional techniques used in abstract writing (see for example (Cleveland and Cleveland, 1983)).

Our method was specified for summarization of one specific type of text: the scientific and technical document. Nevertheless, it is domain independent because the concepts, relations and types of information we use are common across different domains. We believe that our method could eventually be extended to other textual types (i.e., news articles) by specifying a few types of information that allow to compute the topics and expand them according to the readers’ interests. But this issue needs additional exploration.
Chapitre 6

Implementing Selective Analysis in SumUM

In this chapter, we describe SumUM, a computer implementation of Selective Analysis which relies on shallow syntactic and semantic analysis. The system is able to produce short indicative summaries for long technical documents and to expand the summaries with topic elaboration. The system implements the components described in Chapter 5.

6.1 Introduction

SumUM has been implemented in SICStus Prolog (Release 3.7.1) (SICStus, 1998) and Perl (Wall et al., 1996) running on Sun workstations (5.6) and Linux machines (RH 6.0). The following steps have been completed:

- segmentation of the text in main textual units;
- part of speech (POS) tagging of each segment;
- segment and sentence identification;
- shallow sentence parsing and interpretation according to the conceptual model;
- sentence classification according to types of information;
- term extraction and expansion in order to represent the topics;
- pattern matching and template instantiation; and
- content selection and text re-generation.

All the steps but the POS-tagging and the segmentation of the text have been implemented during our research. The POS-tagger was developed by Foster (1991) and the text segmenter was developed by Marco Jacques at Université de Montréal. The POS-tagger is the only external linguistic resource used to implement SumUM.

SumUM relies on traditional string and list processing programs developed by ourselves for the purpose of this research. The information that SumUM produces during the analysis is recorded in the Prolog data base in the form of clauses. The input document is completely analyzed and after that, the information is stored in files in appropriate file directories.
associated to the source document. For an article of 37K characters (13 main sections, 303 sentences and 5004 words), it takes around 3 minutes for segmentation and tagging on a Sun workstations (5.6) and 5 minutes to completely analyze the sentences, extract the information, and produce the abstract on a Linux machines (RH 6.0).

The design of SumUM is modular. Each module can also be executed from a rudimentary user interface. In the following sections, we describe the process that first transforms a string of characters into parsed sentences, then the sentences into a kind of conceptual data base which contains partial syntactic and semantic information, and finally in a summary representation allowing the generation of a text summary.

6.2 Pre-Processing

Pre-processing’s function is to identify the structure of the document. It consists of the following sequential steps (shown in Figure 6.1):

- **Text Segmentation**: The input to SumUM is a plain text without any explicit mark-up. Since several steps in Selective Analysis depend on the structure of the input text, SumUM’s first task is to identify the structural elements of the text: journal information, title, author information, keywords, abstract, acknowledgment, sections and references. SumUM uses typographic features (i.e., non blank lines surrounded by blank lines) and some keywords (e.g., “Keywords”, “Abstract”, “Introduction”, “References” and typical section identifiers “1.”, “2.4”) to identify the end and beginning of a text segment. The segments are stored in separate files with appropriate file names (File.title, File.author, File.section.1, File.references, etc.). The information about the journal, the keywords, and the abstract provided with the source document are only used for evaluation purposes (this will be described in Chapter 7) and do not contribute to the generation of the automatic abstract. The accuracy of this step has been subjectively assessed during experimentation but not formal evaluation has ever been done.

- **POS-tagging**: Each file is passed through the tagger which associates a lexical category to each word. The information produced by the tagger is stored in separate files (File.title.ctag, File.author.ctag, File.section.1.ctag, File.references.-ctag, etc.). The files produced by the POS-tagger contain a record for each word in the input file. The content of each record is either a marker (end of sentence ({sent}), end of paragraph ({para}), end of file ({EOF})) or a word from the text, a sequence of tab characters, the lexical information and the citation form. The output produced by the tagger is shown in Figure 6.2.
6.2. PRE-PROCESSING

![Diagram showing the process from Raw Text to Title Structure](image)

**Fig. 6.1:** Pre-processing and Interpretation in SumUM

**Input Text (File.section.1)**

Introduction

The Automatic Control Department (Instituto de Automática Industrial, Consejo Superior de Investigaciones Científicas, Madrid, Spain) has been developing robotic systems for more than 15 years. This activity...

**POS-tagger Output (File.section.1.ctag)**

```
{para}

Introduction NomC-sing/introduction
The Dete-dart-sgpl-def/a
Automatic AdjQ/automatic
Control NomC-sing/control
Department NomC-sing/department
...
{sent}
This Dete-ddem-sing-def/that
activity NomC-sing/activity
...
{EOF}
```

**Fig. 6.2:** Input Text and POS-Tagging
6.3 Interpretation

The purpose of this step is to obtain a partial syntactic and semantic interpretation of the input text that is based on a set of patterns that account for syntactic and domain specific constructions referring to concepts in the domain. The patterns were identified during the corpus study by the process of tagging all the corpus with POS-tags and then carefully collecting sequences of tags that account for valid linguistic constructions. We used a bi-pos statistical tagger which implements the Viterbi algorithm to find the best sequence of tags for a string of words. The lexical categories used by the POS-tagger are shown in Table 6.1. The categories are combined with other grammatical information such as type of determinant, number, gender, tense, etc., giving a total of 214 different tags.

<table>
<thead>
<tr>
<th>POS</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Det</td>
<td>Determinant</td>
<td>the</td>
</tr>
<tr>
<td>Quan</td>
<td>Quantifier</td>
<td>last</td>
</tr>
<tr>
<td>AdjQ</td>
<td>Adjective</td>
<td>substantial</td>
</tr>
<tr>
<td>Verb</td>
<td>Verb</td>
<td>made</td>
</tr>
<tr>
<td>NomP</td>
<td>Proper Noun</td>
<td>CBR</td>
</tr>
<tr>
<td>Prep</td>
<td>Preposition</td>
<td>within</td>
</tr>
<tr>
<td>Pron</td>
<td>Pronoun</td>
<td>we</td>
</tr>
<tr>
<td>AdvE</td>
<td>Adverb</td>
<td>very</td>
</tr>
<tr>
<td>NomC</td>
<td>Common Noun</td>
<td>retrieval</td>
</tr>
<tr>
<td>ConS</td>
<td>Subordinate conjunction</td>
<td>that</td>
</tr>
<tr>
<td>ConC</td>
<td>Coordinate conjunction</td>
<td>and</td>
</tr>
<tr>
<td>Punc</td>
<td>Punctuation</td>
<td>;</td>
</tr>
<tr>
<td>Affi</td>
<td>Affix</td>
<td>'s</td>
</tr>
<tr>
<td>Ltre</td>
<td>Alphabet</td>
<td>a</td>
</tr>
<tr>
<td>OrdI</td>
<td>Ordinal</td>
<td>fourth</td>
</tr>
<tr>
<td>Post</td>
<td>Post-head particule</td>
<td>a.m.</td>
</tr>
<tr>
<td>Disc</td>
<td>Discourse markers</td>
<td>PS</td>
</tr>
</tbody>
</table>

Tab. 6.1: Part of Speech Categories

To specify linguistic and conceptual patterns, we use sequences of tags from the POS-tagger, our syntactic categories (Table 6.2), words and domain concepts (some patterns are presented in Tables 6.3, 6.4 and 6.5).

The semantic information is organized in a small dictionary that associates lexical items in citation form with semantic tags that we use to represent concepts and relations in the conceptual model. In Table 6.6 we show some of the information in the dictionary. It is implemented with three Prolog predicates: one for each of the POS categories Verb, NomC, and AdjQ.

The overall process of interpretation consists of the following sequential steps: Scanning that identifies sentences and titles in the tagged files; Parsing that identifies syntactic constructions on the sentences relying on the patterns; Semantic that interprets syntac-
6.3. INTERPRETATION

<table>
<thead>
<tr>
<th>POS</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>N+</td>
<td>noun sequence</td>
<td>case base reasoning</td>
</tr>
<tr>
<td>NomP+</td>
<td>proper noun sequence</td>
<td>Aube Billard</td>
</tr>
<tr>
<td>A+</td>
<td>adjective sequence</td>
<td>specific textile</td>
</tr>
<tr>
<td>Adv+</td>
<td>adverb sequence</td>
<td>ever before</td>
</tr>
<tr>
<td>gn</td>
<td>noun group</td>
<td>high-speed parallel</td>
</tr>
<tr>
<td></td>
<td></td>
<td>processing systems</td>
</tr>
<tr>
<td>gv</td>
<td>verb group</td>
<td>will be briefly described</td>
</tr>
<tr>
<td>cue phrase</td>
<td>cue phrase</td>
<td>Additionally</td>
</tr>
</tbody>
</table>

Tab. 6.2: Syntactic Categories

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dete N+</td>
<td>its attention</td>
</tr>
<tr>
<td>A+ N+</td>
<td>industrial robots</td>
</tr>
<tr>
<td>Dete A+ N+</td>
<td>the kinematic chain</td>
</tr>
<tr>
<td>Adv+ A+ N+</td>
<td>very innovative locomotion systems</td>
</tr>
<tr>
<td>NomP+</td>
<td>Instituto de Automatica</td>
</tr>
<tr>
<td>Dete N+ Affi N+</td>
<td>the steam generator’s water chamber</td>
</tr>
<tr>
<td>Quan A+ N+</td>
<td>some successful developments</td>
</tr>
<tr>
<td>Quan N+</td>
<td>most cases</td>
</tr>
<tr>
<td>Verb-PSP N+</td>
<td>related facilities</td>
</tr>
<tr>
<td>Nom+ N+</td>
<td>Motorola processors</td>
</tr>
</tbody>
</table>

Tab. 6.3: Examples of Linguistic Patterns of Noun Groups from the Corpus

Syntactic constructions using the conceptual dictionary; Conceptual Indexing that classifies sentences according to the types of information given in Tables 5.1 and 5.2; and Term Extraction that extracts terms from noun groups and computes term distribution. All the steps are detailed below.

- **Scanning**: SumUM analyses the title, author and reference files first, and then the main sections of the document. SumUM scans the records from tagged files (*.ctag) sequentially and constructs sentence representations which are asserted as Prolog clauses. In this stage common and proper noun distribution is calculated using the information provided by the tagger (words with categories NomC and NomP).

If the record contains a word and its tag, the pair [Structure, Category] is generated. Structure is a Prolog atom representing the word and Category is a list of attribute-value pairs containing the information of the tag. For example, SumUM reads the third record presented in Figure 6.2 and generate the term:

```plaintext
['The',[ (cat,'Dete'), ('Detetype', 'dart'), ('Nbr', 'sgpl'), ('Typ', 'def'), (canon,a) ] ]
```
The order of the elements in this structure is unimportant because the search for particular types of information is based on specific attributes. This kind of structure is used in all levels of analysis of the sentence (lexical, syntactic, and semantic).

Otherwise, the record contains a marker: **SumUM** asserts a clause that represents the sentence (**sentence** in Figure 6.1). It contains the section number, the sentence number (0 for the title) and the pairs so far obtained. Depending on the marker, **SumUM** reads the next record or starts processing the next file.

An example of sentence representation is shown in Figure 6.3.

While **SumUM** constructs sentences from the author and reference files, it creates two lists containing information about proper names that **SumUM** looks for in specific positions of sentences: for example the first author appears in the first line of the author file and the names of the references appear at the beginning of lines in the reference file. These patterns were identified during corpus analysis. This information is subsequently used to interpret proper nouns in sentences from the main sections.

---

**Tab. 6.4:** Examples of Linguistic Patterns of Verb Groups from the Corpus

<table>
<thead>
<tr>
<th>Concept</th>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>author</td>
<td>we</td>
<td>we</td>
</tr>
<tr>
<td>paper</td>
<td>the paper</td>
<td>the paper</td>
</tr>
<tr>
<td>institution</td>
<td>Dete University of gn</td>
<td>The University of Montreal</td>
</tr>
<tr>
<td>section</td>
<td>Section Quan</td>
<td>Section 2</td>
</tr>
<tr>
<td>table</td>
<td>Tables Quan and Quan</td>
<td>Tables 2 and 3</td>
</tr>
<tr>
<td>focus</td>
<td>focus Prep paper</td>
<td>the focus of this paper</td>
</tr>
<tr>
<td>equation</td>
<td>equation (Quan)</td>
<td>equation (3)</td>
</tr>
</tbody>
</table>

**Tab. 6.5:** Examples of Domain Specific Patterns from the Corpus

In this example, the attribute cat contains the lexical category, the attributes DeteType and Type contain information about determinants, and the attribute canon contains the citation form of the word that will be used to compute terms during syntactic analysis.
### 6.3. INTERPRETATION

#### Table 6.6: Examples of the Conceptual Dictionary

<table>
<thead>
<tr>
<th>Conceptual Information</th>
<th>Lexical Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>make known</td>
<td>present, show, identify, report, examine, discuss, introduce, explore, indicate, review, overview, survey, expose, give</td>
</tr>
<tr>
<td>describe</td>
<td>compose, have, consist, make, exhibit, rely on, depend on, comprise, base on</td>
</tr>
<tr>
<td>paper</td>
<td>paper, article</td>
</tr>
<tr>
<td>author</td>
<td>we, I, author</td>
</tr>
<tr>
<td>necessary</td>
<td>necessary, vital, needed, primal, obligatory, mandatory, required</td>
</tr>
<tr>
<td>difficult</td>
<td>difficult, hard, impossible</td>
</tr>
</tbody>
</table>

```
[[Industrial,[(cat,AdjQ),(canon,industrial)]],
[background,[(cat,NomC),(Nbr,sing),(canon,background)]],
[has,[(cat,Verb),(VerbType,aux),(tem,PRES),
(Nbr,sing),(Per,p3),(canon,have)]],
[been,[(cat,Verb),(VerbType,aux),(tem,PSP),(canon,be)]],
[traditionally,[(cat,Adve),(canon,traditionally)]],
[the,[(cat,Dete),(DeteType,dart),(Nbr,sgpl),(Typ,def),(canon,a)]],
[application,[(cat,NomC),(Nbr,sing),(canon,application)]],
[field,[(cat,NomC),(Nbr,sing),(canon,field)]],
[for,[(cat,Prep),(canon,for)]],
[robotics,[(cat,NomC),(Nbr,sing),(canon,robotics)]],
[.,[(cat,Punc),(PuncType,pcst),(canon,.)]]
```

**Figure 6.3: Sentence Representation from the Tagged Files**

- **Parsing**: Sentences and titles are passed through a set of finite state transducers (FST) in order to identify syntactic patterns (noun groups, verb groups, groups of adjectives) and domain specific constructions (i.e., references to the research papers, authors, institutions, bibliographic references, research groups, projects, sections, figures, etc.). This step produces parsed sentences which are also represented as a list of pairs [Structure, Category].

For each pattern, we have implemented a FST that takes as input a sentence representation and looks for a subsequence matching the pattern. Whenever the transducer finds a sequence of elements in the input matching the pattern, it returns a new construction of attribute/value pairs. The matched sequence is replaced by the new construction and the rest of the elements in the input is left unchanged.

The specification of a FST for a pattern $P$ of length $k$ that takes as input a sequence of $n$ elements $S = \bigoplus_{i=1}^{n} S_i$ is as follows:

1. if the input sequence does not contain the pattern, then the output is the input
sequence $S$. This is specified as follows:

$$ \text{FST}_P(S) = S $$

2. otherwise, the input sequence does contain a sequence matching the pattern. Let $i$ be the start position of the first matching sequence in the input. The specification of the FST in this case is:

$$ \text{FST}_P(S) = \bigoplus_{b=1}^{i-1} S_b \oplus \text{construct}_P(\bigoplus_{b=i}^{i+k-1} S_b) \oplus \text{FST}_P(\bigoplus_{b=i+k}^{n} S_b) $$

where $\text{construct}_P(A)$ stands for the parse of $A$ with pattern $P$.

The specification of a FST that looks for non empty sequences of the Cat category (i.e., Cat+) that takes as input a sequence of $n$ elements $S = \bigoplus_{i=1}^{n} S_i$ is as follows:

1. if the input sequence does not contain the Cat category, then the output is the input sequence $S$. The specification is:

$$ \text{FST}_{\text{Cat}+}(S) = S $$

2. otherwise, let $i$ be the position of the first element with Cat category in $S$, and let $k \geq 1$ satisfies the condition : $\forall l : (i \leq l \leq i+k-1)$ : the category of $S_i$ is Cat and the category of $S_{i+k}$ is not Cat or $i+k = n+1$. The specification of the FST is:

$$ \text{FST}_{\text{Cat}+}(S) = \bigoplus_{b=1}^{i-1} S_b \oplus \text{construct}_{\text{Cat}+}(\bigoplus_{b=i}^{i+k-1} S_b) \oplus \text{FST}_{\text{Cat}+}(\bigoplus_{b=i+k}^{n} S_b) $$

where $\text{construct}_{\text{Cat}+}(A)$ stands for the parse of $A$ with pattern Cat+.

The parsing process is the composition of all the transducers in a predefined order. If $\text{FST}_{\text{Pattern}i}$ is to be applied before $\text{FST}_{\text{Pattern}i+1}$, we represent the composition by the expression:

$$ \text{parse}(S) = (\bigcirc_{i=n}^{1} \text{FST}_{\text{Pattern}i})(S) $$

where $\bigcirc$ stands for the composition of finite state transducers. The matching process and the parsing are implemented using Prolog unification.

All the patterns are specified in a grammar that SumUM compiles in order to generate Prolog code implementing the FSTs according to the specification. The time complexity of individual transducer and of their composition is linear on the length of the input.

The order of composition was decided as follows: domain specific transducers need word information so they are applied first. Then, linguistic transducers that detect
more complex syntactic constructions proceed. Among the linguistic transducers, those that analyze Cat+ sequences are applied first and then transducers that look for patterns of fixed length. Those are applied in the following order: if pattern Pattern$_1$ is included in pattern Pattern$_2$ then the transducer FST$_{Pattern_2}$ is applied before that FST$_{Pattern_1}$. Thus, the FST that looks for the construction Dete A+ N+ is applied before that the FST looking for the construction A+ N+.

The parse of a noun group (i.e., output of the construct program) produces the following attribute-value pairs: the original string (attribute string), the canonical form (attribute canon), the syntactic information (attributes syncat, gntype and Nbr), the information about the determinants and quantifiers (attributes Type and Typ), information about anaphoric references (attribute anaphoric), the semantics (attribute sem), the information about adjectives (attribute quality) and information referring to the conceptual model which is optional (attributes conceptual with value dr for domain concept, attribute concept with a value from the conceptual model and other attributes that depend on the particular concept, i.e., id for figures, tables and sections.).

The parse of a verb group produces the following attribute-value pairs: the original string (attribute string), the semantics (attribute pred), the syntactic information (attributes voice, tense, time, type and Nbr), information about adverbial (attribute adv), and the conceptual information which is optional (attributes conceptual with value dr for domain relation and relation with the relation from the conceptual model).

The parse of adjectival and adverbial groups contains the attributes: the original string (attribute string), the citation form (attribute canon), and the optional information from the conceptual model (attribute quality).

The other elements (conjunctions, prepositions, cue phrases, etc.) are left unparsed.

In order to measure the accuracy of the parsing process we manually extracted 303 noun groups and 151 verb groups from abstract found on the (INSPEC, 2000) service. We parsed the abstracts and automatically extracted noun groups and verb groups and we computed recall and precision measures. Recall measures the ratio of the number of correct syntactic constructions identified by the algorithm over the number of correct syntactic constructions. Precision is the ratio of the number of correct syntactic constructions identified by the algorithm over the number of constructions identified by the algorithm. The accuracy of the parser was 86% recall and 86% precision for noun groups and 85% recall and 76% precision for the verb groups. A qualitative analysis will be given in (Saggio and Lapalme, 2000b).

- Semantic: Once all the transducers have been applied, an additional step of analysis is done to interpret noun groups, verb groups and adjectival groups using the informa-
tion provided in the conceptual dictionary.

Noun groups are updated as follows: the citation form of the head of the noun group (attribute \texttt{sem}) is looked up in the conceptual dictionary and its semantic tag, if found, is inserted in the construction (attributes \texttt{conceptual} and \texttt{concept}). Information about the adjectival modifiers of the noun group is also extracted from the dictionary and placed in the representation (attributes \texttt{quality}).

For the interpretation of verb groups, the citation form (value of the attribute \texttt{pred}) is looked up in the conceptual dictionary and its semantic tag (if one exists) is included in the representation (attributes \texttt{conceptual} and \texttt{relation}). In case of ambiguity (i.e., the verb “present” can be used to mark the topic of a document and to describe an entity), all the tags found on the dictionary for a particular lexical item are used to update the construction.

Adjectives in adjectival groups are looked up in the dictionary and the group is updated with their associated semantic tags in the same way.

After that, domain specific transducers that combine simple concepts into more complex ones are applied, for example in order to interpret expressions such as the goal of the X project.

In Figure 6.4, we schematically present the analysis of complete sentences, while in Figure 6.5 we show examples of parsed sentence fragments: the input sequence, the matched pattern and the attribute-value pairs obtained during the parsing (more examples can be found on Appendix F).

- **Conceptual Indexing**: Its function is to classify sentences according to types of information as specified in Tables 5.1 and 5.2, this step is essential for both indicative and informative selection. \texttt{SUMUM} examines the conceptual attributes (\texttt{concept}, \texttt{quality} and \texttt{relation}) and updates the conceptual index which specifies to which type of information the sentence could eventually contribute. \texttt{SUMUM} verifies the co-occurrence of semantic tags in order to classify the sentence: for example in order to classify a sentence as \textbf{Topic of Section} it will verify the co-occurrence of the concept \texttt{paper component} and a relation \texttt{make known}. The conceptual index is organized in two structures: the \textit{indicative index} for the indicative sentences and \textit{informative index} for the informative ones. These are Prolog clauses that record sentences positions and their types.

- **Term Extraction**: Finally, \texttt{SUMUM} extracts terms and acronyms and their expansions from the sentence. In order to extract the terms \texttt{SUMUM} looks for the values of the attributes \texttt{canon} and \texttt{sem} in each noun group in the sentence. The value of the attribute \texttt{canon} is a sequence of words (in citation form) and the value of the attri-
Fig. 6.4: Syntactic and Conceptual Information inParsed Sentences

But the semantic representation contains the title of the document, the information about the authors, the titles of each section and the sentences syntactically and semantically analyzed. Each element is asserted as Prolog a fact containing the parsed sentence and its position. Additionally, information about the structure of the document (section order and paragraph structure of each section) is generated. This implementation of **SumUM** ignores other aspects of the structure of a text like lists and enumerations, and most of the process overlooks the information about the paragraph structure. Nevertheless, in future improvements of **SumUM** that will be taken in consideration.

The accuracy of the parsing process has not been addressed during this research, but a
This activity (Dete N+)
(sem,activity), (conceptual,dc), (concept,activity),
anaphoric,definite), (synct,gn), (gntype,GN2),
(string,[This,activity]), (canon,[activity]), (DeteType,ddem),
(Typ,def), (Nbr,plur)

The Automatic Control Department (Dete A+ N+)
(sem,department), (conceptual,dc), (concept,institution), (synct,gn),
gntype,GN4), (string,[The, Automatic, Control, Department]),
(canon,[automatic, control, department]), (DeteType,dart), (Typ,def),
(Nbr,sing)

the long-term goal (Dete A+ N+)
(sem,goal), (conceptual,dc), (concept,goal), (synct,gn),
gntype,GN4), (string,[the, long-term, goal]), (canon,[long-term,
goal]), (DeteType,dart), (Typ,def), (Nbr,sing)

Inove (1994) (NomP (Year))
(synct,gn), (gntype,ref), (reftype,'Ref1'), (conceptual,dc),
(concept,reference), (researcher,'Inove'), (year,1994),
(string,['Inove','(','1994,')]}, (Nbr,nil)

briefly outlines (Adv Verb)
(conceptual,dr), (pred,outline), (type,none), (relation,'to make
know'), (synct,gv), (Nbr,plur), (tense,sim_pre), (voice,active),
(time,pres), (string,[briefly, outlines]), (canon,outline),
(adv,briefly)

FIG. 6.5: Attribute Pair Values in Parsed Sentence Fragments

sound methodology of evaluation for the parser component will be proposed and applied in
future work.

6.4 Indicative Selection

The purpose of this step is to instantiate indicative templates with sentences classified
as indicatives by the previous step, the conceptual index is used for this purpose. The
process is shown in Figure 6.7.

Templates are implemented with the Prolog unary predicate template. The argument
is a list whose first element is the type of the template and the rest of the list is a set of
attribute-value pairs which we represent with the Prolog term \texttt{Slot:Filler}. Where, \texttt{Slot} is an atom and \texttt{Filler} is an atomic or a list depending on the slot (for the specification of the templates refer to Section 5.4).

\textbf{SumUM} looks for indicative sentences appearing on the \textit{indicative index} and proceeds with the analysis. \textbf{SumUM} excludes from the analysis sentences matching one of the following conditions: (i) it contains cue phrases such as \textit{however, although, for example}, etc. at the beginning of the sentence (we used a list of cue phrases from Marcu (1997b)), (ii) it contains potential dangling anaphora: \textit{the first X, the following X}, etc. (iii) it starts with lower case or with punctuation; (iv) it ends with an element different from a full stop; (v) it contains mathematical or quantity concepts; (vi) it contains bibliographic references (this is not required for the types of information \textbf{Situation, Problem Identification, Need} and \textbf{Problem Solution}). Two procedures are applied to extract information from sentences and to instantiate templates: the first one was designed to look for specific patterns, the second one is based on the identification of domain relations and the extraction of arguments. Both are detailed below.

### 6.4.1 Extraction using Patterns

Given a sentence $S$ and a type of information $T$, \textbf{SumUM} applies an specific Prolog program to interpret sentence $S$ according to the type of information $T$. Basically, it consists of the verification of some linguistic-conceptual patterns and the instantiation of a template.
Patterns are represented with the predicate `pattern`. Its arguments are the name of the pattern and the pattern itself. The pattern itself is a list of pairs. The first element of the pair is either the atom `skip` (it indicates to overlook zero or more elements on the input) or a list of attribute-value pairs that stands for syntactic and conceptual information to be verified during the matching process using Prolog unification. The second element of the pair is a Prolog variable. Informative patterns also include an argument for the specific topic under consideration which also appears on the list as value for the attribute `canon` that is used to record terms obtained from noun groups (this implements a kind of dynamic pattern).
6.4. INDICATIVE SELECTION

<table>
<thead>
<tr>
<th>Type</th>
<th>Pattern Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signaling Structural</td>
<td>$SKIP_1 + \text{structural} + SKIP_2 + \text{show graphically} + ARGUMENT + \text{eos}$</td>
</tr>
<tr>
<td>Signaling Concept</td>
<td>$SKIP_1 + \text{overview} + \text{Prep} + ARGUMENT + \text{eos}$</td>
</tr>
<tr>
<td>Topic</td>
<td>$\text{gn} + \text{author} + \text{make known} + \text{Prep} + \text{research paper} + DESCRIPTION + \text{eos}$</td>
</tr>
<tr>
<td>Author’s Goal</td>
<td>$SKIP_1 + \text{goal of author} + \text{define} + GOAL + \text{eos}$</td>
</tr>
<tr>
<td>Goal of $TOPIC$</td>
<td>$SKIP + \text{goal} + \text{Prep} + TOPIC + \text{define} + GOAL + \text{eos}$</td>
</tr>
<tr>
<td>Definition of $TOPIC$</td>
<td>$SKIP + TOPIC + \text{define} + \text{gn}$</td>
</tr>
<tr>
<td>Elaboration of $TOPIC$</td>
<td>$SKIP + TOPIC + \text{elaborate}$</td>
</tr>
</tbody>
</table>

Prolog Implementation

```
pattern(goal_of_author(1), [[skip,X1],
[(concept,goal_of_author)],X2], [(relation,'to define')],X3),
[skip,X4], [('PunctType',pcst)],X5]).
```

```
pattern(definition(1),TERM, [[skip,X1], [[canon,TERM]],X2],
[(relation,'to define'), (voice,active)],X3),
[(syncat,gn)],X4)).
```

Figure 6.8: Specification of Indicative and Informative Sentence Patterns and Prolog Representation of the Author’s Goal and Definition

In Figure 6.8, we show patterns and their Prolog implementation (Signaling Structural, Signaling Concept, Topic and Author’s Goal are indicative patterns, and Goal, Definition and Elaboration of topic are informative patterns). \text{eos} means end of sentence.

Each element of the pattern matches one or more elements of the sentence: conceptual, syntactic, and lexical elements match one element while variables match zero or more (they are represented by the atom skip). One program scans the sentence verifying the occurrence of the conceptual and linguistic constructions (this is done unifying the attribute-values pairs of the pattern with those from the sentence) and instantiates the variables of the pattern with parsed fragments: for a pattern with \( n \) elements, the program will return a list with \( n \) instantiated Prolog variables.
For example, the following sentence matches the pattern **Author’s Goal**:

\[
\text{goal of author define}
\]

Our aim is to devote the human skills to more complex and intellectual parts of the operation, i.e., detection and vehicle driving, which do require effort, while the robot will perform the hardest and more precise jobs, localisation and harvesting.

during the matching process five variables \((x_i)\) get instantiated with the following elements: (i) \(x_1\) with the empty list (because the goal of author is the first element of the sentence), (ii) \(x_2\) with the element matching the goal of author concept, i.e., Our aim, (iii) \(x_3\) with the element matching the define relation, i.e., is, (iv) \(x_4\) with the parsed fragment to the right of the define relation, i.e., to devote..., and (v) \(x_5\) with the full stop. These variables are used to instantiate a template of type goal of author (note that this template is the implementation of the **Conceptual Goal** given in Section 5.4.1, page 77, Figure 5.12).

The goal of author template contains the following slots: (i) id to be filled with a unique integer identifier automatically obtained, (ii) marker to be filled with the parsed concept goal of author, (iii) pred to be filled in with the predicate matching the define relation, (iv) goal to be filled in with the parsed fragment to the right of the relation, and (v) sentence instantiated with the sentence position.

The topic candidates are the terms extracted from the slot goal and are recorded also as Prolog database clauses along with their relevance, which is the sum of the relevance of the terms as specified in Section 5.4.1. The predicate template tbl is used to record the information about the terms and the weight. The template gets instantiated as shown in Figure 6.9.

Note that a sentence could match different patterns associated to a type of information \(T\), in that case **SumUM** selects the longest pattern of type \(T\) that matches the sentence.

### 6.4.2 Extraction using Domain Relations

When the type of information \(T\) requires the extraction of a domain relation \(R\) and its arguments (i.e., types Topic of Document, Topic of Section, Possible Topic, Author Interest, Author Development, etc.) the procedure is as follows:

1. **SumUM** looks for an instance of the domain relation \(R\) matching a set of restrictions and creates two windows around it. This produces a structure of the form \([\text{Left Window, Domain Relation, Right Window}]\) where the first and third components are parsed fragments (possibly empty) and the second component is the parsed domain relation;

2. **SumUM** uses grammatical information to decide where to look for the arguments. If the value of the attribute voice of the Domain Relation is active, the main argument is
uninstantiated template

template([goal_of_author, id:ID, marker:MARKER, pred:PRED, goal:GOAL, sentence:[SEC,FID]])

instantiate template

template([goal_of_author, id:81, marker:[...,[(Concept,goal_of_author),(Sem,aim),(conceptual,dc), (concept,goal), (syncat,gn), (gntype,'GN2'), (string,['Our',aim]), (canon,[aim]), ('DeteType',dpos), ('Typ',def), ('Per',p1)]], pred:[be,[is],active,sim_pre,pre,sing], goal:[[to,[Cat,'Prep'],(..., ...),sentence:[1,24]])

topic candidates and weight

template_tbl(goal_of_author:81:[[complex, and, intellectual, part], [detection], [effort], [hard, and, precise, job], [human, skill], [localisation], [operation], [robot], [vehicle, driving]]:[1,24]:121)

Fig. 6.9: Author’s Goal Instantiated during Indicative Selection

extracted from the Right Window with a specific Prolog program while the grammatical subject is taken from the Left Window. Otherwise, the grammatical subject is the atom nil and the argument is extracted from the Left Window.

The sentence is segmented in positions where specific patterns occur. These include eos (end of sentence), ConC gv (conjunction followed by verb), cue phrase gv (cue phrase followed by verb), Semi_colon (semi colon), PunC paper component (punctuation followed by paper component).

Each domain relation dictates the concepts that can act as arguments (for example, the filler of the slot Who of the make known relation can be one of the following: author, research_paper, study, research, and work). In order to extract the main argument, a specific procedure is used which looks for patterns of noun groups such as gn CONT* where CONT is (gn), (Prep gn), (ConC gn), (gVpos_par gn), (gVpar_pre Prep gn), (gVpre_pre gn), (comma gn) among others. For the domain relation most to the right of the sentence and for the argument of topical information, all the parsed string on the right window is extracted. If the voice is passive, before extracting the argument some elements from the beginning of
the window will be ignored like sentence fragments starting by prepositions and cue phrases.

For example, the sentence:

\[
\text{We first describe these four visual code representations and then discuss the interaction techniques for manipulating them.}
\]

is first segmented according to the two domain verbs and then the windows are defined around them:

- **Left Window**: \text{make known} \ \
- **Right Window**: \text{first describe these four visual code representations}

- **Left Window**: \text{make known} \ \
- **Right Window**: \text{then discuss the interaction techniques for manipulating them}

In this case, as both verbs are in the active voice, **SumUM** extracts the arguments from the right window to instantiate the **what** slot, and the grammatical subject from the left window to instantiate the **who** slot. The slot **where** is instantiated with nil because there is no indication of the concept **research paper**. The instantiated templates are shown in Figure 6.10.

Note that a sentence could be used to instantiate different templates, but in **SumUM** if a sentence was used to instantiate templates of type **Topic of Document** or **Topic of Section** or **Entity Introduction** or **Entity Identification** or **Signaling Structures** or **Signaling Concepts**, that sentence will not be considered to instantiate other types of templates.

```
template([topic,id:7,pred:[describe,[first,describe]],active,
sim_pre,pre,nil], who:[author_paper,..., where:nil,what: [[[these,...(string,[these,four,visual,
code,representations])], (canon,[visual,code,representation])]],[
sentence:[1,20]])).

template([topic,id:8,pred:[discuss,[then,discuss]],active,sim_pre,
pre,nil], who:nil, where:nil, what: [[[the,...,(gntype,'GN*'),(string,[the,interaction,techniques])
... ([symcat,gv)...(string,[manipulating]),(canon,manipulate)]],
[them,[(cat,'Pron'),...],sentence:[1,20]])).
```

**Fig. 6.10**: **Topic of the Document** Instantiated during Indicative Selection
6.4.3 Selecting the Content

The process of content selection is straightforward. **SumUM** looks for a match between each element of the list *topical structure* and the terms on the topic candidates that are contained in the clauses *template.tbl*. **SumUM** verifies the substring relation between Prolog lists and records the information about the match in another list, the *matched structure*, which contains a pair for each matched term: the term and the information about the templates matching the term. This process takes time proportional to the size of the topical structure times the size of the union of the topic candidates. In a practical situation, the size of the topical structure is small compared with the number of terms of the document.

Next, **SumUM** selects one template for each element on the *matched structure*, ordering them by weight. If there are more than one, it uses heuristics about the type of information and the position found on this list (Section 5.7). The *indicative content* is a list containing the identification of the templates selected by this process. The terms from the *template.tbl* as well as their expansions are recorded on the list *potential topics*. The expansions are found on the acronym information, and by transforming the AVL tree into a list (this is done with built-in Prolog predicates from the Association List package) and retrieving all the terms sharing the semantics of the topic candidates. A pretty print of the tree is shown in Figure 6.11; it contains the semantics of the terms, the terms themselves, and the positions where they occur.

At the end of this process **SumUM** has computed two structures: the list of templates *indicative content* that will be used by the generation step to produce the indicative abstract, and the list of terms *potential topics* that will be used by the informative selection step to look for informative sentences that expand those terms.

6.5 Informative Selection

The purpose of this step is to look for sentences expanding the potential topics that will be used to instantiate the informative templates described in Section 5.4.2. The informative selection process is shown in Figure 6.12.

**SumUM** retrieves from the *term tree* the sentence positions where the potential topic occurs and verifies if the sentence appears in the *informative index*. The sentence is retrieved from the *text representation* and both the sentence and the potential topic are passed through a process of pattern matching. If the sentence matches a pattern of the specific type, a template will be instantiated and the potential topic becomes a topic of the document which is appended to the list of topics so far obtained and asserted on the data base. The patterns **SumUM** uses are dynamic in the sense that they contain some fixed concepts, relations and syntactic constructions and a specific position for the potential topic under consideration.
(Agribot, [Agribot])-[[[2,0],[2,1],[2,2]],...  
...  
(condition, [condition])-[[[6,17],[7,7]],...  
(condition, [laboratory, condition])-[[[1,26]],...  
(condition, [light, condition])-[[[1,22]],1]  
(condition, [real, condition])-[[[10,6]],1]  
(condition, [special, condition])-[[[2,13]],1]  
...  
(operation, [acceleration, operation])-[[[5,25]],1]  
(operation, [cutting, operation])-[[[6,12]],1]  
(operation, [detach, operation])-[[[10,1]],1]  
(operation, [fast, operation])-[[[5,44]],1]  
(operation, [grasping, operation])-[[[6,2]],1]  
(operation, [high, velocity, operation])-[[[11,6]],1]  
(operation, [manual, operation])-[[[3,5]],1]  
(operation, [operation])-[[[1,17]],...  
(operation, [relate, operation])-[[[1,5]],1]  
(operation, [simultaneous, two-arm, operation])-[[[10,3]],1]  
(operator, [human, operator])-[[[2,5],[4,1],[8,15],[11,1]],4]  
(operator, [operator])-[[[2,6],[2,7],[2,10],[4,3]],...  

**Fig. 6.11:** Information on the Term Tree

For example, the following sentence:

```
potential topic

The NTRS provides the following services to the user. elaborate
```

matches the informative pattern of topic elaboration presented in Figure 6.8, the three variables X_i instantiated with: (i) the empty list, (ii) the parsed potential topic, and with (iii) the parsed elaborate relation. A template of type elaboration is instantiated and asserted in the database.

The informative templates contain the slot Content which is instantiated with the sentence avoiding structural elements (figures, tables, etc.) and the slot Topic which is instantiated with the term. The other components are the type of template, its identifier and the sentence position.

### 6.6 Generation

The purpose of this step is to generate the indicative abstract from the indicative content and the list of topics. The reader uses the list of topics in order to obtain additional
information from the informative templates (i.e., the informative abstract).

**SumUM** arranges the information according to the conceptual order specified and merges the “topical” templates implementing the algorithm discussed in 5.7. The templates are stored on the list text plan.

The generation is straightforward: one generation procedure is called for each template on the text plan implementing the schemas presented in Section 5.7. First, **SumUM** produces a list with the words. It then prints the list with the correct punctuation.

Three main processes update the list of words:

1. the reformulation of domain verbs. The parameters passed to the program are the position where the verb is to be generated (beginning or middle of the sentence), the voice (active or passive), the number (singular or plural), and the tense. These parameters account for the differences: *Presents vs. presents, Presents vs. is presented,*
Presenting the Indicative Abstract...

The complex problem of inspection and maintenance of the steam generator in nuclear power plants was approached using previously gained expertise and, as a result, an innovative solution was achieved with the development of two co-operative robots, remotely controlled from a tele-operation station incorporating tele-presence. The department was leading a project for the introduction of a robotics system whose mission was to avoid human operators having to enter the steam generator's water chamber. Proposes an innovative solution to the serious problem of inspection and maintenance of the steam generator tubes, which must be checked regularly to detect leakage of radioactive water from the primary to the secondary circuit. Another interesting activity was the realization, within the framework of a EUREKA project, of a tele-manipulator for servicing a new concept of urban infrastructures. Shows the RIMHO walking robot and ROBUR arm exchanging gas filter in IUI (Industrializable urban infrastructures) demonstration.

Abstract WC : 149

Identified Topics : [IUI] [RIMHO] [chamber] [climb,robot] [control,station] [control,system] [different,robotic,system] [human,operator] [industrial,robot] [infrastructure] [robot] [robot,arm] [system] [water,chamber] [wheel,mobile,robot] [whole,control,system]

Fig. 6.13: Output of SumUM for the article ‘Climbing, walking and intervention robots”, Industrial Robot, Vol 24 Issue 2, 1997

is presented vs. are presented and studied vs. studies. The Prolog program returns the appropriate form found on a table of verbs. The reformulation is inserted on the list.

2. the reformulation of domain concepts, author related entities, and noun groups introduced by a demonstrative. A table of concepts and its pre-defined linguistic realization is consulted using the domain concept and the position in the sentence. The possessive pronoun of the author related expression and the demonstrative in the case of a demonstrative noun group are replaced by the determinant “the” (or “The”). The expression is appended to the list.

3. the reformulation of parsed fragments. For lexical elements (preposition, conjunctions, etc.) and punctuation, the original word is extracted from the structure and put on the list. For verb groups, adjectival and adverbial groups, and noun groups unrelated to the conceptual model, the value of the attribute string is retrieved from the structure. The case of the first word is amended according to the position and the expression is appended to the list. For an acronym, the expansion is retrieved and appended to the list surrounded by parenthesis.

The generation of the merged template is simple: it uses the previous procedures and inserts the markers semicolon and conjunction (and, and also) between propositions and arguments.

The list is used to produce a string with appropriate punctuation (i.e., word, word; word : word) (word word, etc.) and to generate a word count. The sentences, the word count and the list of topics are printed as shown in Figure 6.13, the input article is shown in Figure 6.14.

For the informative abstract two options are available: the presentation of all the informative sentences or the consultation of some of them according to the reader’s interests:
Climbing, walking and intervention robots
Manuel Armada

[...] 

Introduction

The complex problem of inspection and maintenance of the steam generator in nuclear power plants was approached using previously gained expertise and, as a result, an innovative solution was achieved with the development of two co-operative robots, remotely controlled from a tele-operation station incorporating tele-presence (appears as first sentence of the automatic abstract) [...] 

Another interesting activity, also reported here, was the realization, within the framework of a EUREKA project, of a tele-manipulator for servicing a new concept of urban infrastructures (appears as fourth sentence of the automatic abstract) [...] 

Nuclear power plant steam generator inspection and maintenance robot - SIROIN [...] 

Our department was leading a project (1986-91) for the introduction of a robotics system whose mission was to avoid human operators having to enter the steam generator’s water chamber (appears as second sentence of the automatic abstract) [...] 

With SIROIN we have proposed an innovative solution to the serious problem of inspection and maintenance of the steam generator tubes, which must be checked regularly to detect leakage of radioactive water from the primary to the secondary circuit (appears as third sentence of the automatic abstract) [...] 

Self-propelling climbing robot

The self-propelling climbing robot [2] is a simple prototype of three [...] 

The RIMHO walking robot

The RIMHO walking robot is a prototype developed with the aim of studying the potential possibilities [...] 

With such geometry the machine can walk with a clearance of 350mm (see Figure 4 The RIMHO walking robot ) (appeard in the last sentence of the automatic abstract) [...] 

An excellent demonstration (see Figure 6 ROBUR arm exchanging gas filter in IUI demonstration ) was held and the IUI system was (appears in the last sentence of the automatic abstract) [...] 

Conclusions

An overview of different robotic systems developed for performing tasks in hazardous environments has been presented. The research efforts in this field of robotics are continuing in [...] 

1. Should the reader want information about all the topics, SumUM will retrieve all the informative templates from the informative data base, will sort the templates according to the position slot, and will present the slot Content using the reformulation of parsed fragment sentences using the parameter beginning of sentence. SumUM avoids the generation of the same information twice (this is done controlling the information on the Position slot).

2. If the reader chooses some of the topics, only the templates with value of Topic matching the chosen topics are retrieved and presented.

6.7 Limitations of the Implementation

SumUM was implemented in order to demonstrate of viability of Selective Analysis. We used state-of-the-art techniques in natural language processing including noun and verb group identification and conceptual tagging because we wanted to produce a robust domain independent text summarization system. In this perspective, we decided to implement most of the process with reliable finite state techniques (Roche and Schabes, 1997). While finite state techniques behaved well in the task of identification of concepts and relations of our conceptual model, they are not the ultimate solution to modeling natural language. Context-free formalisms seem far more appropriate to deal with the real complexities of natural language, though they were not necessary for this research.

We have not addressed here the question of text understanding: SumUM is able to produce text summaries but it is not able to demonstrate intelligent behavior (answering questions, paraphrasing, anaphora resolution, etc.). Nevertheless, in future research we want to address topic elaboration based on questions directly formulated by the reader (i.e., who did X? what is X? which is the relation between X and Y? who previously worked with X?). This will require the incorporation of natural language understanding techniques which will allow improvement of the text summaries: SumUM will be able to generate in some cases anaphoric references or will be able to replace an anaphora by its antecedent. A partial solution to that problem will help in the identification of additional information about the “topics” computed by SumUM. In fact, given a text like “(1) The paper presents two text summarization systems... (2) The first is based on a conceptual model... (3) The second uses a language model to ...” SumUM is unable to interpret the expressions “The first” and “The second” as anaphoric references to “two text summarization systems” and so it is unable to elaborate the topic “text summarization system” using sentences (2) and (3).

SumUM relies on the output produced by a shallow text segmenter and on a statistical POS-tagger, which incorporates some noise to the process. The segmentation algorithm is based on the structure of electronic documents found on the Emerald Electronic Library, and so it has to be adapted for other text structures. No rigorous tests have ever been done to measure the accuracy of the POS-tagger, but the success rate is probably around 96%\(^2\) meaning that if the average sentence length is 25 tokens, then the tagger produces on the

\(^2\)George Foster, personal communication.
average an error per sentence. We do not have any procedure to automatically "correct" the output of the tagger (if the tagger interprets the word "reports" in the sentence "The paper reports experiments on automatic abstracting" as a noun instead of a verb, then SumUM will overlook the useful information of that sentence). SumUM does not deal with ambiguities like those that occur in expressions that help us identify concepts in our domain, so SumUM is unable to answer the following questions: Is the expression "the paper" a reference to the research article or to "a pliable substance made usually of vegetable matter and used to write or print on"? or Is the expression "we" a reference to the author of the paper or to a group that includes the author of the paper?

The patterns incorporated in our grammar account for only a few linguistic constructions (174 indicative patterns and 87 informative patterns) and our conceptual dictionary contains only a few words (241 domain verbs, 163 domain nouns and 129 adjectives). While the verbs and nouns we use to identify the information for the indicative abstract are more likely to appear in the technical domain, the verbs and adjectives that we use in order to identify the information for the informative abstract are general in nature and portable to other domains. We have only implemented the following types of indicative information: Situation, Problem, Solution, Need, Topic of Document, Topic Description, Possible Topic, Topic of Section, Signaling, Entity Introduction, Entity Identification, Development, Study, Interest, Goal, Focus, Inference, Method, Result, Experiment, Summary. The informative types implemented are: Definition, Description, Elaboration, Goal, Focus, Identification, Need, Interest, Relevance, Advantage, Positiveness, Practicality, Effectiveness, Development. It would be interesting to incorporate techniques to semi automatically learn patterns and lexical items from tagged corpus as in the experiments described by Lehnert et al. (1992). But instead of using a matching process between instantiated templates and source documents we would like to try a match between human abstracts and source documents.

Our prototype only analyses sentences for the specific purpose of text summarization and implements some patterns of generation observed in the corpus. Nevertheless, the representation produced by SumUM is rich enough to allow further analysis, for example by incorporating partial syntactic and rhetorical parsing. The latter not only will help in the extraction of arguments and the deletion of sentence fragments but also in the task of expanding the text with additional information. In fact, if a sentence is presented that elaborates a topic, probably the context of the sentence is also important for the topic. This could also be addressed taking into consideration the structure of the document which is calculated by SumUM. Other aspects of the text structure will be addressed in the future such as the identification of lists and enumerations. Despite these limitations, SumUM is able to produce quite acceptable abstracts as will be shown in Chapter 7.

The summaries produced by the system were extensively evaluated. Though, the implementation was only assessed during experimentation. Our future work will address the design of a sound methodology of evaluation of the different components: this will be done by comparing the output of each component with an accurate analysis produced by human
informants. For example, given a document, we will compare the accuracy of the terms computed by \textbf{SumUM} by matching the automatic terms with the correct terms of the document. This will be used to calculate recall and precision measures for comparison purposes as presented in page 121. The very same method could be applied for the identification of the types of information, the partial parsing of the sentences and the extraction of arguments.

### 6.8 Summary

This chapter described \textbf{SumUM}, a prototype of Selective Analysis which goes all the way through from a raw text to a summary. This is accomplished by the processes of text segmentation, POS-tagging, partial syntactic and semantic analysis, sentence classification, template instantiation, content selection, text regeneration and topic elaboration. \textbf{SumUM} was developed in order to validate this new methodology of text summarization.

The system was implemented in Prolog and relies on classical algorithms of string processing and pattern matching. While \textbf{SumUM} is an experimental system, it has been tested in long documents from several sources and domains including Industrial Robot Journal, Internet Research Journal, Assembly Automation Journal, Information Management & Computer Security Journal, COMPEL Journal, Computer Journal and BioInformatics (some automatic abstracts are included in Appendix J). The results so far obtained are already good. Nevertheless, improvements have to be done to transform the actual implementation in a production environment.
Chapitre 7

Evaluating Content and Text Quality in Selective Analysis

Evaluation has become a central issue in Natural Language Processing. In this chapter we describe three evaluations methodologies in which we have compared different summarization technologies in order to assess the degree of effectiveness of our summarization method.

7.1 Introduction

The quality and success of human produced abstracts have already been addressed in the literature (Grant, 1992; Kaplan et al., 1994; Gibson, 1993) using linguistic criteria such as cohesion and coherence, thematic structure, sentence structure, and lexical density. But in automatic text summarization, this is an emergent research topic.

Content evaluation assesses if the automatic system is able to identify the intended “topics” of the source document. Text quality evaluation assesses the readability, grammar and coherence of the summary. The evaluations can be done in intrinsic or extrinsic fashions as defined by Spark Jones and Galliers (1995).

An intrinsic evaluation measures the quality of the summary itself. This is done by comparing the summary with the source document by measuring how many “main” ideas of the source document are covered by the abstract or by comparing the content of the automatic summary with an ideal abstract (gold standard) produced by a human (Cole, 1995).

An extrinsic evaluation measures how helpful a summary is in the completion of a given task. For example, given a document which contains the answers to some predefined questions, readers are asked to answer those questions using the abstract. If the reader correctly answers the questions, the abstract is considered of good quality for the given question-answering task. Variables measured can be the number of correct answers and time to complete the task. Recent experiments (Jing et al., 1998) have shown how different parameters such as the length of the abstract can affect the outcome of the evaluation.
We have evaluated the abstracts produced by **SumUM** in three different situations:

- The first evaluation, which was intrinsic, addressed the issue of indicativeness: how **SumUM** performs in indicating the essential content of the source document using as gold standard the abstract published with the source document. Here, we measured how many of the “topics” were covered by the automatic abstract. We also evaluated how adequate the sentences automatically generated were when compared with human sentences. For sentence acceptability we relied on human assessment.

- The second evaluation, which was extrinsic, was repeated in three opportunities with different materials. It addressed the indicativeness of the text in a categorization task using descriptors published with the source document. In this evaluation we also assessed the quality of the whole text. For both aspects, we relied on human judgment. Text content was evaluated extrinsically.

- Finally, the third evaluation, which was intrinsic, addressed the issue of informativeness: how **SumUM** performed in the task of selecting “interesting” informative sentences, using as ideal abstracts sets of sentences selected by typical readers. Here, we measured the coverage of important information.

In the three evaluations we used other available summarization methodologies for comparison purposes. In this chapter we will give the details of the three evaluation methodologies and the results so far obtained.

### 7.2 First Evaluation: Human Abstracts as Ideal Abstracts

The most extensive independent evaluation of automatic summarization systems to date was done in the context of the TIPSTER SUMMAC evaluation (Mani et al., 1998). That evaluation was extrinsic because it addressed the utility of automatic abstracts for completing a given task (categorization, question-answering, etc.). However, even though this effort created a corpus of full-texts for evaluation purposes, there is a clear lack of resources for the evaluation of technical articles. We addressed this by constructing our own evaluation resources with technical articles published in electronic journals on the Web. We use as gold standard for evaluation the abstracts published with the source documents (as was the case in (Lin and Hovy, 1997) but for different purposes) and we compared the terms appearing in the automatic abstracts with the terms appearing in the abstracts provided by the journal. We do not compare sentences with sentences because the abstracts published together with source documents usually contain sentences difficult to match with those of the source document (Teufel and Moens, 1998).

The performance of **SumUM** was measured relative to two other summarization methodologies: abstracts produced using **Word Distribution**, and abstracts produced using the commercially available **Microsoft '97 Summarizer**. We implemented the **Word Distribution** method by computing the distribution of nouns (common nouns and proper nouns) in all the text (using the result of the POS-tagging process and the canonic form
7.2. **FIRST EVALUATION: HUMAN ABSTRACTS AS IDEAL ABSTRACTS**  

of the words) and then by associating a score to each sentence (the sum of the distribution of its nouns). To produce the abstract, the method chooses top ranked sentences until the desired compression rate is achieved. This technique, though simple, has been used alone or in combination with other methods in order to produce summaries (Luhn, 1958; Brandow et al., 1995; Kupiec et al., 1995). The commercial system used here for comparison purposes has already been used in evaluation of text summarization systems (see for example (Marcu, 1997a; Barzilay and Elhadad, 1997)).

7.2.1 Experiment 1

For this experiment, we used 25 technical articles found on the Emerald electronic library (Industrial Robot Journal, Internet Research Journal, Assembly Automation Journal, Information Management & Computer Security Journal, and COMPEL Journal) and on the electronic version of the Computer Journal. The articles contain titles, author identification, a short list of keywords (2 to 5), an indication of the type of article (technical, case study, etc.), an abstract, the text of the article, and references. The articles are quite long (from 13K characters to 36K characters with an average of 23K characters). The given abstracts and lists of keywords were not considered in order to produce the automatic abstracts.

We automatically extracted a list of terms from the given abstracts using the resources described in chapter 6 (finite state transducers and term extraction), considering only those terms appearing in the abstract and in the source document. We produced abstracts using **SumUM** and extracted the list of topics from the abstracts: we considered terms appearing in the indicative abstracts and in informative sentences\(^1\). We computed the compression ratio in number of words\(^2\) for the automatic abstract (CSA) and the abstract provided by the journal (CAA)\(^3\). Except for one document, **SumUM** always produced more verbose abstracts than the provided abstract. The compression ratio (between 91% and 96% with an average of 94.4%) was always greater than the compression ratio of 90% used in other summarization evaluations (Mani et al., 1998).

Next, we produced two additional abstracts for each document: one by **Word Distribution** and other using the **Microsoft'97 Summarizer**, the compression ratio being the smaller of CSA and CAA (i.e., allowing the other abstracts to be at least as verbose as **SumUM**). In order to produce the abstract by word distribution, we used the results from the pre-interpretation step in **SumUM**. In order to produce the abstract with **Microsoft'97 Summarizer**, we had to format the source document in order for the **Microsoft'97 Summarizer** to be able to recognize the structure of the document (titles, sections, paragraphs and sentences).

\(^1\)When this experiment was carried out, the process of term expansion described in page 91 has not yet been implemented.

\(^2\)Compression is given by the following formula \(100 \times (1 - \frac{\text{# words in the abstract}}{\text{# words in source document}})\).

\(^3\)For this experiment we do not use character compression because **Microsoft'97 Summarizer** works based on word compression.
Following this, we extracted terms from the abstract obtained by word distribution and from the abstract obtained using Microsoft’97 Summarizer. We used the very same techniques than in selective analysis (i.e., we interpreted the sentences in both abstracts identifying noun groups and extracting terms). To sum up, for each document in the test set we have the following lists:

- \( Terms(AA, i) \) denotes the list of terms extracted from the ideal abstract of document \( i \) and represents our gold standard for this evaluation;
- \( Terms(SA, i) \) denotes the list of terms extracted from the abstract by SumUM for document \( i \);
- \( Terms(MS, i) \) denotes the list of terms extracted from the abstract by Microsoft’97 Summarizer for document \( i \); and
- \( Terms(WD, i) \) denotes the list of terms extracted from the abstract by word distribution for document \( i \).

The terms in the lists \( Terms(Method, i) \) where \( Method = SA, WD \) or \( MS \) were compared with the terms in the list \( Terms(AA, i) \) and recall, precision and F-score measures were calculated for the three methodologies and each individual source document. Recall, precision and F-score are standard measures used in the evaluation of Information Retrieval and Information Extraction systems (Cole, 1995). Here we adapt these measures as follows:

- Recall measures the ratio of the number of relevant terms identified by the automatic method (i.e., terms appearing in both the list of terms of the automatic abstract and in the list of terms of the ideal abstract) over the number of the relevant terms in the author abstract. It was computed using the following formula:

\[
Recall(Terms(Method, i)) = \frac{||Terms(AA, i) \cap Terms(Method, i)||}{||Terms(AA, i)||}
\]

- Precision is the ratio of relevant terms identified by the automatic method (i.e., terms appearing in the list of terms of the automatic abstract and in the list of terms of the ideal abstract) over the number of terms identified by the automatic method. It was computed using the following formula:

\[
Precision(Terms(Method, i)) = \frac{||Terms(AA, i) \cap Terms(Method, i)||}{||Terms(Method, i)||}
\]

- F-score is a composite score that combines the precision and recall measures. It was computed using the following formula:

\[
Fscore(Terms(Method, i)) = \frac{2 \cdot Recall(Terms(Method, i)) \cdot Precision(Terms(Method, i))}{Recall(Terms(Method, i)) + Precision(Terms(Method, i))}
\]

As an illustration, in Figures 7.1 and 7.2, we show the four abstracts for a document in the evaluation and in Table 7.1, we show the terms extracted from those abstract, the topics correctly identified by each method (in bold), and the three measures.
### Abstract Published by the Journal

This work presents a robot prototype designed and built for a new aided fruit-harvesting strategy in highly unstructured environments, involving human-machine task distribution. The operator drives the robotic harvester and performs the detection of fruits by means of a laser range-finder, the computer performs the precise location of the fruits, computes adequate picking sequences and controls the motion of all the mechanical components (picking arm and gripper-cutter). Throughout this work, the specific design of every module of the robotized fruit harvester is presented. The harvester has been built and laboratory tests with artificial trees were conducted to check range-finder's localization accuracy and dependence on external conditions, harvesting arm's velocity, positioning accuracy and repeatability; and gripper-cutter performance. Results show excellent range-finder and harvesting arm operation, while a bottleneck is detected in gripper-cutter performance. Some figures showing overall performance are given.

### Abstract by Microsoft

Design and implementation of an aided fruit-harvesting robot (Agribot). Detection and localisation of fruits; and The pointing and picking processes are always executed simultaneously, that is, while the operator is pointing to the fruits, the computer generates optimal picking sequences and the manipulator harvests the fruits. Design of the robotic system. Fruit localisation module. The operator detects the fruit and uses the joystick to place the laser spot on the fruit. The spherical coordinate, indicating the location of the fruit, is stored in a dynamic data structure where the robot looks for target fruits pending being harvested. Mechanical design of the picking arm. Grasping and detaching tool motors of the localisation module and picking arm. Localisation results. Harvester arm performance. Detaching results. The presented work follows a new approach with remarkable advantages concerning fruit localisation and harvester mechanical structure. Figure 1 Configuration of the robotic fruit harvester Agribot. Figure 2 Statistical model of harvesting environment: (a) fruit height distribution; (b) fruit depth distribution; (c) fruit-picking zone. Figure 3 Parallelogram structure of Agribot's picking arm. Figure 6 Schematic representation of the laser range-finder testing setup. Table II Velocity and acceleration figures of Agribots picking arm. Figure 8 Specification of testing picking trajectories. Plate 1 View of the robotic fruit harvester Agribot during laboratory tests.

**Fig. 7.1:** Abstracts used for evaluation purposes for the Document "Design and implementation of an aided fruit-harvesting robot (Agribot)", Industrial Robot, Vol 25 Issue 5, 1999 (continues on Figure 7.2)
Abstract by Word Distribution

Associated strategies, such as centering the fruit in the image during the approximation movements toward the fruit, or stereoscopic vision with triangulation techniques from the matching of the images from two cameras for fruit location have been implemented. In doing so, the operator uses a joystick that allows orientation of a two-degree of freedom pan / tilt mechanism which carries the laser telemeter and is placed on the cabin (see Figure 1 Configuration of the robotic fruit harvester Agribot). All the different modules comprising the robotic system, that is, the detection and localisation module, the picking arm and detaching tool, are integrated by the control and processing computer which carries out the general functions of data acquisition (encoders, IR, pressure sensors, telemeter, inductive limit switches, joystick and devices of the control board), data processing (filtering, location calculations, grasping sequence, trajectories determination, testing and others) and control of external devices (motor references, electrical brakes, board display, saw, pneumatic valve). Concerning the fruit localisation system, we have carried out an evaluation of the quality of the measure, that is standard deviation of the range measurement, as a function of external lighting conditions, angle of incidence of the laser beam on to the surface to be measured, distance from the target, color and reflectance of the target. Nevertheless, a pointer coaxial to the laser beam has been added to the system so that an approximate idea of the laser beam direction is obtained (see Figure 1 Configuration of the robotic fruit harvester Agribot).

Abstract by SumUM

Almost all the proposed prototypes rely on a human guided vehicle for solving the first task, while detection and localisation of fruits, which appear to be the most difficult problem, are faced in an automatic mode of operation based on artificial vision. Presents the mechanical and electronic design of the robot harvester including all subsystems, namely, fruit localisation module, harvesting arm and gripper-cutter as well as the integration of subsystems. The harvester has been tested in laboratory conditions: tests are described and results are given together with some conclusions of the work. Presents the specific mechanical design of the picking arm addressing the reduction of undesirable dynamic effects during high velocity operation. The Agribot’s approach presents a semi-automatic way of operation, with realistic goals, combining harmoniously the human and machine functions. The harvesting strategy that inspires the robotic harvester relies on an operator that will guide the vehicle in the grove and, once stopped, detects the fruits, while the robotic system locates them, plans the picking sequence and makes the approximation and detaching of the fruit. Shows schematic view of the detaching tool and operation and view of the robotic fruit harvester Agribot during laboratory tests.

Fig. 7.2: Abstracts used for evaluation purposes for the Document “Design and implementation of an aided fruit-harvesting robot (Agribot)”, Industrial Robot, Vol 25 Issue 5, 1999
<table>
<thead>
<tr>
<th>Source</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAA</td>
<td>accuracy; computer; dependence; detection; fruit; harvester; laboratory test; motion; operator; picking arm; repeatability; result; robotic harvester; sequence; specific design; unstructured environment; and work.</td>
</tr>
<tr>
<td>TSA R. .53 P. .22 F. .31</td>
<td>Agribot; every target; arm; condition; design; detaching tool; detection; difficult problem; dynamic; fruit; fruit localisation module; function; groove; harvester; human guided vehicle; integration; laboratory; laboratory condition; laboratory test; laser telemetry; localisation; operation; operator; picking arm; picking sequence; problem; result; robot; robotic fruit harvester; robotic harvester; robotic system; schematic view; system; task; test; tool; vehicle; velocity; way; and work.</td>
</tr>
<tr>
<td>TWD R. .35 P. .09 F. .14</td>
<td>Agribot; IR; angle; approximate idea; approximation movement; associate strategy; board display; cabin; color; computer; condition; configuration; control; control board; data acquisition; data processing; detection; device; different module; distance; electrical brake; encoders; evaluation; external device; freedom pan; fruit; fruit localisation system; fruit location; function; general function; image; incidence; inductive limit switch; joystick; laser beam; laser beam direction; laser telermeter; localisation module; location calculation; measure; motor reference; operator; orientation; picking arm; pneumatic valve; pointer coaxial; pressure sensors; quality; range measurement; robotic fruit harvester; robotic system; see; sequence; standard deviation; stereoscopic vision; surface; system; target; testing; tilt mechanism; tool; trajectory determination; triangulation technique; two camera; and two-degree.</td>
</tr>
<tr>
<td>TM R. .47 P. .16 F. .16</td>
<td>Agribot; Agribot picking arm; Localisation result harvester arm performance; acceleration figure; aided fruit-harvesting robot; arm; computer; configuration; design; detaching tool; detection; dynamic data structure; environment; fruit; fruit depth distribution; fruit height distribution; fruit localisation; fruit-picking zone; harvester mechanical structure; implementation; joystick; laboratory test; laser range-finder testing setup; laser spot; localisation; localisation module; location; manipulator; mechanical design; new approach; operator; parallelogram structure; picking arm; pointung and picking processes; presented work; remarkable advantage; result; robot; robotic fruit harvester; robotic system fruit localisation module; schematic representation; sequence; specification; spherical coordinate; statistical model; target fruit; testing; tool motor; trajectory; velocity; and view.</td>
</tr>
</tbody>
</table>

Tab. 7.1: Terms Extracted from the four Abstracts and Recall (R), Precision (P) and F-score (F)

### 7.2.2 Results of Indicativeness

In Table 7.2, we present the figures obtained for the 25 articles and the three methodologies and the average information. These numbers indicate that **SumUM** performs better than word distribution and **Microsoft '97 Summarizer** when the source document is a
technical article and the compression ratio is high (more than 91%). There is indication that *SumUM* performs better in precision with a gain of 125% over the two other methods, performing better than the other two methods in 20 cases. This is due to the fact that the terms produced by our method are those additionally elaborated in the source document, and not only “mentioned” in the indicative abstract; as a consequence the original list gets reduced. Although the average recall for the 25 articles indicates a gain of 25% over word distribution and *Microsoft'97 Summarizer*, there is no clear indication of better performance in general (*SumUM* performed better than the other methods in 10 cases, word distribution in 5 cases and *Microsoft'97 Summarizer* in 7 cases). Regarding F-score, we have obtained a gain of 85% over word distribution and a gain of 84% over *Microsoft'97 Summarizer*.

<table>
<thead>
<tr>
<th>Article Number</th>
<th>SumUM</th>
<th>Word Distribution</th>
<th>Microsoft</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.29</td>
<td>.25</td>
<td>.27</td>
</tr>
<tr>
<td>2</td>
<td>.27</td>
<td>.36</td>
<td>.31</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>.50</td>
<td>.25</td>
<td>.33</td>
</tr>
<tr>
<td>5</td>
<td>.30</td>
<td>.23</td>
<td>.26</td>
</tr>
<tr>
<td>6</td>
<td>.33</td>
<td>.18</td>
<td>.23</td>
</tr>
<tr>
<td>7</td>
<td>.40</td>
<td>.36</td>
<td>.38</td>
</tr>
<tr>
<td>8</td>
<td>.14</td>
<td>.08</td>
<td>.10</td>
</tr>
<tr>
<td>9</td>
<td>.53</td>
<td>.22</td>
<td>.31</td>
</tr>
<tr>
<td>10</td>
<td>.40</td>
<td>.09</td>
<td>.15</td>
</tr>
<tr>
<td>11</td>
<td>.25</td>
<td>.06</td>
<td>.10</td>
</tr>
<tr>
<td>12</td>
<td>.27</td>
<td>.09</td>
<td>.13</td>
</tr>
<tr>
<td>13</td>
<td>.19</td>
<td>.20</td>
<td>.19</td>
</tr>
<tr>
<td>14</td>
<td>.37</td>
<td>.35</td>
<td>.36</td>
</tr>
<tr>
<td>15</td>
<td>.50</td>
<td>.15</td>
<td>.23</td>
</tr>
<tr>
<td>16</td>
<td>.11</td>
<td>.07</td>
<td>.09</td>
</tr>
<tr>
<td>17</td>
<td>.25</td>
<td>.09</td>
<td>.13</td>
</tr>
<tr>
<td>18</td>
<td>.29</td>
<td>.21</td>
<td>.24</td>
</tr>
<tr>
<td>19</td>
<td>.09</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>20</td>
<td>.27</td>
<td>.40</td>
<td>.32</td>
</tr>
<tr>
<td>21</td>
<td>.29</td>
<td>.18</td>
<td>.22</td>
</tr>
<tr>
<td>22</td>
<td>.13</td>
<td>.13</td>
<td>.13</td>
</tr>
<tr>
<td>23</td>
<td>.29</td>
<td>.24</td>
<td>.26</td>
</tr>
<tr>
<td>24</td>
<td>.33</td>
<td>.25</td>
<td>.25</td>
</tr>
<tr>
<td>25</td>
<td>.57</td>
<td>.20</td>
<td>.20</td>
</tr>
</tbody>
</table>

**Tab. 7.2:** Detailed Recall, Precision and F-score for the 25 Technical Articles and Average Information across Documents
7.2. FIRST EVALUATION : HUMAN ABSTRACTS AS IDEAL ABSTRACTS

7.2.3 Influence of the Term Extraction Algorithm in the Evaluation

In order to have a clearer picture about the effect of using our program of term extraction for evaluation, we decided to repeat the previous experiment but this time comparing the nouns (common nouns and proper nouns in citation form) appearing in automatic abstracts with nouns appearing in the abstracts provided by the journal (gold standards). In this evaluation, in addition to SumUM, Microsoft’97 Summarizer, and Word Distribution, we considered the following simple methods of producing abstracts:

- Introduction Method (Intro) : this method chooses sentences (in positional order) from the introduction of the document until the desired compression rate is achieved;
- Begin of Paragraph & Noun Distribution Method (Para & Word Distribution) : this is the Word Distribution method applied to sentences at begin of paragraph : top ranked sentences are chosen until the desired compression rate is achieved;
- Introduction & Noun Distribution Method (Intro & Word Distribution) : this is the Word Distribution method applied to sentences in the introduction of the article : top ranked sentences are chosen until the desired compression rate is achieved.

For this evaluation we used 95 technical articles from the Emerald Electronic Library. For each article and method, we produced an abstract at the same compression that SumUM had produced. Then, we extracted the nouns from each abstract using the information provided by the POS-tagger. We also tagged the author abstract and extracted the list of nouns from it. To sum up, we have the following sets of nouns (1 ≤ i ≤ 95):

- Nouns(AA, i) denotes the set of nouns extracted from the ideal abstract of document i and represents our gold standard for this evaluation;
- Nouns(Method, i) denotes the set of nouns extracted from the abstract produced by the automatic method (SumUM, Microsoft, Word Distribution, Intro, Intro & Word Distribution, Para & Word Distribution) for document i.

We computed recall, precision and F-score for all the methods and abstracts as follows:

\[
\text{Recall}(\text{Nouns}(\text{Method, } i)) = \frac{|\text{Nouns}(\text{AA, } i) \cap \text{Nouns}(\text{Method, } i)|}{|\text{Nouns}(\text{AA, } i)|}
\]

\[
\text{Precision}(\text{Nouns}(\text{Method, } i)) = \frac{|\text{Nouns}(\text{AA, } i) \cap \text{Nouns}(\text{Method, } i)|}{|\text{Nouns}(\text{Method, } i)|}
\]

\[
\text{Fscore}(\text{Nouns}(\text{Method, } i)) = \frac{2 \times \text{Recall}(\text{Nouns}(\text{Method, } i)) \times \text{Precision}(\text{Nouns}(\text{Method, } i))}{\text{Recall}(\text{Nouns}(\text{Method, } i)) + \text{Precision}(\text{Nouns}(\text{Method, } i))}
\]

The average information over the 95 documents is presented in Table 7.3. Even if in this experiment, SumUM obtained the better score and the ranking is the same as before, the numbers greatly differ from the previous evaluation revealing that our term extraction algorithm gave advantage to SumUM over the other methodologies, as a consequence, this ask for the construction of gold standards and measures of similarity not biased to any particular method. This little experiment helped us to corroborate that a very simple method of choosing sentences from introduction combined with noun distribution produced very good results for our test set.
### Table 7.3: Average Recall, Precision and F-score over 95 Documents

<table>
<thead>
<tr>
<th>System</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumUM</td>
<td>.32</td>
<td>.36</td>
<td>.32</td>
</tr>
<tr>
<td>Intro &amp; Word Distribution</td>
<td>.30</td>
<td>.35</td>
<td>.31</td>
</tr>
<tr>
<td>Microsoft</td>
<td>.28</td>
<td>.35</td>
<td>.29</td>
</tr>
<tr>
<td>Intro</td>
<td>.22</td>
<td>.27</td>
<td>.23</td>
</tr>
<tr>
<td>Para &amp; Word Distribution</td>
<td>.23</td>
<td>.28</td>
<td>.23</td>
</tr>
<tr>
<td>Word Distribution</td>
<td>.22</td>
<td>.26</td>
<td>.22</td>
</tr>
</tbody>
</table>

### 7.2.4 Content-Based Measures of Evaluation

We have also explored the use of content-based measures of similarity such as the vector space model for the evaluation of text summarization systems as recently proposed by Donaway et al. (2000). In this model (Salton, 1988), each text or abstract is represented as a vector of term weights. In our case, we represented the automatic abstracts in the evaluation as vectors of nouns weights normalized with inverse document frequency (Salton, 1988), the ideal abstracts were also represented as vectors of nouns weights. The similarity between an automatic abstract and the ideal abstract is given by the cosine of the angle between the vectors representing the abstracts. If the cosine is close to 1, both abstracts can be deemed as similar in content. In this evaluation we have obtained good results for SumUM when compared with the methods used in the previous evaluation. The average cosine for each method computed over 95 documents is presented in Table 7.4. Details about this experiment will be given in (Saggion and Lapalme, 2000b).

### Table 7.4: Average Cosine over 95 Documents

<table>
<thead>
<tr>
<th>System</th>
<th>Cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumUM</td>
<td>.46</td>
</tr>
<tr>
<td>Microsoft</td>
<td>.43</td>
</tr>
<tr>
<td>Intro &amp; Word Distribution</td>
<td>.38</td>
</tr>
<tr>
<td>Para &amp; Word Distribution</td>
<td>.32</td>
</tr>
<tr>
<td>Word Distribution</td>
<td>.31</td>
</tr>
<tr>
<td>Intro</td>
<td>.27</td>
</tr>
</tbody>
</table>

### 7.2.5 Acceptability

In the experiment we describe in the next section, we only address the issue of sentence acceptability. In order to evaluate the acceptability of the sentences produced using our method, we used human judges and we asked them to decide if the sentences produced by SumUM are acceptable to be included in indicative abstracts when compared with human produced sentences.
7.2.6 Experiment 2

In this experiment, we relied on 3 human judges with experience in reading technical articles. We presented the judges with a list of 150 randomly selected sentences from three different sources: (1) 50 sentences written by professional abstractors (PA), (2) 50 sentences written by the authors of source documents (SD) which contain the information reported in the professional abstracts of our corpus or in the abstracts we generate, and (3) 50 sentences produced by our system (SA) (only “topical” sentences). In Table 7.5, we show one sentence of each type. The sentences were presented in random order and without source indication. We asked the judges to decide for each sentence if it was acceptable or not to be included in indicative abstracts. The sentences had to be judged independent one another. As in the evaluation done by Coch (1996), we give the judges some criteria for sentence acceptability, such as “good grammar” and “correct spelling”, and also a short statement “the sentences are generally brief, and usually, don’t contain references to the source document.” We gave the judges two examples of good indicative abstracts written by professional abstractors. The judges were informed that they could consider a sentence acceptable even if it contained minor errors. They were also told that some acronyms could appear without expansion and that this situation was also acceptable (this will be an issue once we evaluate text acceptability). We used the vote of the majority in order to consider a sentence as acceptable.

<table>
<thead>
<tr>
<th>SA</th>
<th>Presents the architecture of the agent; describes its design and implementation; and gives some examples showing the cluster labels generated by the clustering algorithm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA</td>
<td>Presents a more efficient Distributed Breadth-First Search algorithm for an asynchronous communication network.</td>
</tr>
<tr>
<td>SD</td>
<td>The software presented in this article adds new motion features to the Aria-Delta parallel robot.</td>
</tr>
</tbody>
</table>

Tab. 7.5: Sentences from the 3 Sources: SumUM (SA), Professional Abstractor (PA), and Source Document (SD)

7.2.7 Result of Acceptability

The results of the experiment are presented in Table 7.6. These indicate a good acceptability rate for SumUM when compared with human generated sentences, Most of the sentences automatically generated were unacceptable for the very same reasons that human produced sentences were unacceptable (too brief, too long, use of passive voice, impersonal, etc.). The sentences produced by professional abstractors were always more acceptable than the other two types of sentences. Note that the information from the source documents comes from sentences that are sometimes too informative or that sometimes include titles and captions; this in part explains the results.
7.3 Second Evaluation: Extrinsic

Our second experiment which was extrinsic addressed the evaluation of content and text quality, this time using human judges. We compared the abstracts produced by SumUM with abstracts produced by Microsoft’97 Summarizer and with abstracts published with source documents (usually author abstracts). As in the SUMMAC evaluation (Mani et al., 1998), we performed a categorization task. We presented judges with abstracts and five lists of keywords (descriptors). The judges had to decide to which list of keywords the abstract belongs given the fact that different lists share some keywords. Those lists were obtained from the journals where the source documents were published. The idea behind this evaluation is to see if the abstract conveys the very essential content of the source document in order to help readers to complete a classification task.

In order to evaluate the quality of the text, we asked the judges to provide an acceptability score between 0 and 5 for the abstract (0 for unacceptable and 5 for acceptable) based on the following criteria taken from Rowley (1982) which were only suggestions and were not compulsory:

- good spelling and grammar;
- clear indication of the topic of the source document (topical sentence);
- impersonal;
- single paragraph;
- conciseness (no unnecessary references to the source document);
- readable and understandable;
- acronyms are presented along with their expansions; and
- other criteria that the judge considered important as an experienced reader of abstracts of technical documents.

We told the judges that we would consider the abstracts with scores above 2.5 as acceptable. Some criteria are more important than others; for example, evaluators do not care about impersonal or personal style but care about readability.
7.3.1 Evaluation at Université de Montréal

Materials

Source Documents: we used ten source documents from the journal “Industrial Robots” found on the Emerald Electronic Library (all technical articles). The articles were downloaded in plain text format. The documents are quite long texts (between 15K and 41K characters) containing 3834 words on the average (with a minimum of 2481 and a maximum of 6196 excluding punctuation), and an average of 180 sentences (with a minimum of 89 and a maximum of 288).

Abstracts: We produced ten abstracts with SumUM and computed the compression ratio in number of words, then we produced ten abstracts with Microsoft’97 Summarizer using a compression rate at least as high as SumUM (i.e., if SumUM produced an abstract representing 3.3% of the source, we produced the Microsoft’97 Summarizer abstract representing 4% of the source). We extracted the ten abstracts and the ten lists of keywords published with the source documents. To sum up, we had 30 different abstracts and ten list of keywords. Examples are shown in Figures 7.3, 7.4, and 7.5.

The first prototype telexistence master slave system for remote manipulation experiments was designed and developed. Telexistence can be divided into two categories: telexistence in the real environment that exists at a distance, and is connected via a robot to the place where the user is located; and telexistence in a virtual environment that does not exist was created by a computer. Shows HRP (Humanoid Robotics Project). Covers the realization of R-Cubed (R3 Real-time Remote Robotics).

Fig. 7.3: SumUM abstract for the document “Telexistence and R-Cubed”, Industrial Robots 26(3) used for evaluation purposes

Abstract Characteristics: the abstracts by SumUM were produced considering the following types of information: Topic, Possible Topic, Topic of Section, Entity Introduction, Study, Goal, Focus, Development and Signaling Structural. The abstracts we obtained by Microsoft’97 Summarizer sometimes contained titles, captioning and bibliographic references making the text difficult to understand but not to classify. We believe that the abstracts published by the journal are good texts with impersonal as well as personal style.

Keyword Characteristics: the keywords used in this evaluation are presented in Table 7.7. 90% of the author abstracts explicitly mention at least one discriminant keyword in their descriptors (i.e., different from “robots”). Some descriptors could be deduced from terms

---

4We had to format the source document in order for the Microsoft’97 Summarizer to be able to recognize the structure of the document (titles, sections, paragraphs and sentences).
Telexistence and R-Cubed
Augmented telexistence
Figure 3 Diagram of an augmented telexistence system shows the schematic diagram of the augmented telexistence system and Figure 4 Virtual Telesar at work shows the virtual telexistence anthropomorphic robot used in the experiment [4].
Humanoid Robotics Project : HRP
Figure 2 Telesar (telexistence surrogate anthropomorphic robot) developed. Figure 3 Diagram of an augmented telexistence system. Figure 4 Virtual Telesar at work. Figure 9 HRP (Humanoid Robotics Project) general plan.

**Fig. 7.4:** Microsoft’97 Summarizer abstract for the Document “Telexistence and R-Cubed”, Industrial Robots 26(3) used for evaluation purposes

Telexistence (tele-existence) is technology which enables a human being to have a real time sensation of being at a remote location, while giving the person the ability to interact with the remote and/or virtual environments. He or she can “telexist” (tele-exist) in a real environment where the robot exists or in a virtual environment that a computer has generated. It is also possible to telexist in a mixed environment of real and virtual, which is called augmented telexistence. The concept of telexistence, i.e. virtual existence in a remote or computer-generated environment, has developed into a national R&D scheme called R-Cubed (Real-time Remote Robotics). Based on the scheme the National R&D Project of “Humanoid and Human Friendly Robotics”, Humanoid Robotics Project (HRP) in short, was launched in April 1998. This is an effort to integrate telerobotics, network technology and virtual reality into networked telexistence.

**Fig. 7.5:** Abstracts published with the source document “Telexistence and R-Cubed”, Industrial Robots 26(3) used for evaluation purposes

appearing in the author abstract.

**Evaluation Forms**: We arranged the 30 abstracts and list of keywords in forms in order to satisfy the following:

- each judge evaluates six abstracts of six different documents; and
- each abstract is evaluated by three different judges.

In order to do so, we produced 5 different forms, each form contained six different abstracts randomly\(^5\) chosen which are versions of six different documents. Each abstract was printed as a paragraph in a different page without indication of the method used in order to produce the abstract. Each page included the following:

---

\(^5\)Random numbers for this evaluation were produced using software provided by SICSTus Prolog.
### Tab. 7.7: Keywords from Industrial Robots used in the Evaluation at Université de Montréal

- an abstract;
- 5 lists of keywords;
- a field to be completed with the quality score associated to the abstract; and
- a field to be filled with comments about the abstract.

One of the lists of keywords was the one published with the source document, the other four were randomly selected from the set of 9 remaining keyword lists, they were printed in the form in random order. One page was also available to be completed with comments about the task, in particular with the time it took to the judges to complete the evaluation. We produced three copies of each form for a total of 15 forms.

### Subjects

We had a total of 15 human judges or evaluators. Our evaluators were two professors and 13 students of second year of the M.Sc. program in Information Science of École de Bibliothéconomie et des Sciences de l’Information (EBSI) at Université de Montréal. All of the subjects had reading and comprehension skills in English. This group was chosen because they have knowledge about what constitutes a good abstract and they are educated to become professionals of Information Science.

### Evaluation Procedure

The evaluation was performed in one session of one hour at the EBSI. Each human judge received a form (so he/she evaluated six different abstracts) and an instruction booklet. No other material was required for the evaluation (i.e., dictionary). We asked the judges to read carefully the abstract. They had to decide which was the list of keywords that matched the abstract (they could chose more than one or none at all) and then, they had to associate

<table>
<thead>
<tr>
<th>Neural networks, Programming, Robots</th>
<th>Robots, Food industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer networks, Conferencing, Internet, Virtual reality</td>
<td>Language, Robots</td>
</tr>
<tr>
<td>CAD/CAM, Manufacturing, Simulation, VR</td>
<td>Control, Language, Robots</td>
</tr>
<tr>
<td>Deburring, Fettling, Force sensors, Robots</td>
<td>Robots, Grippers, Assembly</td>
</tr>
<tr>
<td>Robots, Service, Telexistence, Teleoperation, VR</td>
<td>Robots, Teleoperator</td>
</tr>
</tbody>
</table>
a numeric score to the abstract representing its quality based on the given criteria. This procedure produced three different evaluations of content and text quality for each of the 30 abstracts. The overall evaluation was completed in a maximum of 30 minutes.

Results

For each abstract, we computed the average quality using the scores given by the three judges (i.e., sum of the scores divided by three). We considered that the abstract indicated the essential content of the source document if two or more judges (i.e., the majority) were able to chose the correct list of keywords for the abstract.

<table>
<thead>
<tr>
<th>Article</th>
<th>Microsoft Abstract</th>
<th>SumUM</th>
<th>Author Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Indicative</td>
<td>Quality</td>
<td>Indicative</td>
</tr>
<tr>
<td>1</td>
<td>yes</td>
<td>2.83</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>no</td>
<td>1.50</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>no</td>
<td>1.33</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>yes</td>
<td>0.50</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>yes</td>
<td>2.00</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>yes</td>
<td>1.70</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>yes</td>
<td>1.33</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>yes</td>
<td>1.33</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>yes</td>
<td>0.33</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>yes</td>
<td>1.83</td>
<td>no</td>
</tr>
<tr>
<td>Average</td>
<td>80%</td>
<td>1.46</td>
<td>80%</td>
</tr>
</tbody>
</table>

Tab. 7.8: Results of Human Judgment about Indicativeness and Text Quality. Data from the evaluation carried out at École de Bibliothéconomie et des Sciences de l’Information de l’Université de Montréal

The results for individual articles and the average information are shown in Table 7.8. For a given source document and type of abstract, the value in column ‘Indicative’ contains the value ‘yes’ if the majority of the evaluators have chosen the source document list of keywords for the abstract and ‘no’ on the contrary. The value in column ‘Quality’ is the average acceptability for the abstract.

Content: In all the cases, the abstracts published with the source documents were correctly classified by the evaluators. In contrast, the automatic abstracts were correctly classified in 80% of the cases. It is worth noting that the automatic abstracts were in all cases less verbose that the abstracts published with the source documents and that the automatic systems did not use the journal abstracts nor the lists of keywords or the information about the journal.

Quality: The figures for text acceptability indicate that the abstracts produced by Microsoft’97 Summarizer are below the acceptability level of 2.5, because in general titles and captioning are present in the abstracts making it difficult to read. The abstracts produced
by SumUM are above the acceptability level of 2.5, while the human abstracts are highly acceptable. This clearly indicates that SumUM produces quite good abstracts, sometimes even as acceptable that the human produced ones (article 2, 3 and 8). Nevertheless, the low score obtained by SumUM is due to problems in the order of the information, incompleteness in content, sentences too long, dangling anaphora, and style too technical. This information was obtained from the page reserved for comments.

### 7.3.2 Evaluation at McGill University

The objectives and procedures for this evaluation are the same as those presented in Section 7.3.1, but we used different materials and subjects.

**Materials**

**Source Documents**: we used twelve source documents from the journal “Industrial Robots” found on the Emerald Electronic Library (all technical articles). The articles were down-loaded in plain text format. These documents are quite long texts with an average of 23K characters (minimum of 11K characters and a maximum of 41K characters). They contain an average of 3472 words (minimum of 1756 words and a maximum of 6196 words excluding punctuation), and an average of 154 sentences (with a minimum of 85 and a maximum of 288).

**Abstracts**: we produced twelve abstracts using our method and computed the compression ratio in number of words, then we produced twelve abstracts by Microsoft'97 Summarizer using the same compression ratio as our method. We extracted the twelve abstract and the twelve lists of keywords published with the source documents. We had a total of 36 different abstracts and ten lists of keywords.

**Keyword Characteristics**: the descriptors used in this evaluation are presented in Table 7.9. 70% of the author abstract explicitly mention at least one discriminant keyword in its descriptor (i.e., different from “robots”).

**Evaluation Forms**: We arranged the 36 abstracts and list of keywords in forms in order to satisfy the following:

- each judge evaluates six abstract of six different documents; and
- each abstract is evaluated by three different judges.

In order to do so, we produced 6 different forms, each form contained six different abstracts randomly chosen which are versions of six different documents. Each abstract was printed as a paragraph in a different page. Each page included the following:

- an abstract;
- 5 lists of keywords;

---

6 Two list of keywords were repeated.
Control, Language, Robots  
Robots, Health care  
Robots, Service, Telexistence, Teleoperation, VR, Virtual Reality  
Metrology, Photogrammetry, Robots  
Language, Neural networks, Robots  
Calibration, Force sensing, Robots, Torque sensing  
Robots, Welding, Automotive  
Robots, Teleoperator  
Asbestos, Insulation, Robotics  
Robotics

Tab. 7.9: Keywords from Industrial Robots used on the Evaluation at McGill University

- a field to be completed with the quality score associated to the abstract; and
- a field to be filled with comments about the abstract.

One of the lists of keywords was the one published with the source document, the other four were randomly selected from the set of 11 remaining keyword lists, they were printed in the form in random order. One page was also available to be completed with comments about the task, in particular with the time it took to the judges to complete the evaluation. We produced three copies of each form for a total of 18 forms.

<table>
<thead>
<tr>
<th>Article</th>
<th>Microsoft Abstract</th>
<th>SumUM</th>
<th>Author Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>yes 2.66</td>
<td>Indicative Quality 2.93 yes 4.16</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>no 1.36</td>
<td>yes 3.66 yes 4.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>no 1.16</td>
<td>no 3.00 no 4.06</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>yes 3.00</td>
<td>yes 4.00 yes 4.33</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>no 2.16</td>
<td>no 1.76 yes 4.00</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>yes 2.16</td>
<td>yes 4.00 no 4.53</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>no 0.83</td>
<td>yes 2.50 yes 4.40</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>yes 2.33</td>
<td>yes 3.00 yes 4.00</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>yes 2.16</td>
<td>no 2.66 yes 3.66</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>yes 2.16</td>
<td>yes 4.00 yes 3.31</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>yes 2.40</td>
<td>no 2.70 no 4.26</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>no 1.16</td>
<td>no 3.33 no 4.00</td>
<td></td>
</tr>
</tbody>
</table>

Average 70% 1.98 70% 3.15 80% 4.04

Tab. 7.10: Results of Human Judgment about Indicativeness and Text Quality. Data from the evaluation carried out at McGill Graduate School of Library & Information Studies
Subjects

We had a total of 18 human judges or evaluators. Our evaluators were 18 students of the M.Sc. program in Information Science at McGill Graduate School of Library & Information Studies. All of the subjects had good reading and comprehension skills in English.

Evaluation Procedure

The evaluation was performed in one hour session at McGill University. Each human judge received a form (so he/she evaluated six different abstracts) and an instruction booklet. No other material was required for the evaluation (i.e., dictionary). The procedure was the same as in the previous evaluation. We had three different evaluations of content and text quality for each of the 36 abstracts. The overall evaluation was completed in a maximum of 40 minutes (some materials used in this evaluation are included in the Appendixes G, H and I). For each abstract, we computed the same measures than in the previous evaluation (see page 154). The results for individual articles and the average information are shown in Table 7.10.

Results

Content: In 80% of the cases, the abstracts published with the source documents were correctly classified by the evaluators. In contrast, the automatic abstracts were correctly classified in 70% of the cases.

Quality: The figures about text acceptability indicate that the abstracts produced by Microsoft'97 Summarizer are below the acceptability level of 2.5. The very same reasons than in the previous experiment were given by the evaluators. The abstracts produced by our method are above the acceptability level of 2.5, and the human abstracts are highly acceptable (one time SumUM was better than the human abstract). This time the critics to SumUM were bad flow of ideas, too long sentences, dangling anaphora, style too technical, sentences without grammatical subject, and poor cohesion.

7.3.3 Evaluation at John Abbott College

This evaluation was similar to the two previous evaluations, although here we considered only automatic abstracts. The summarization systems included in this evaluation were: SumUM, Extractor and n-STEIN. Extractor (Turney, 1999) takes a text file as input (plain ASCII text, HTML, or e-mail) and generates a list of keywords and keyphrases as output. It produces between 3 and 30 phrases. On average, it generates the requested number of phrases, but the actual number for any given document may be slightly below or above the requested number, depending mainly on the length of the input document. We used Extractor 5.1 which is distributed for demonstration. We downloaded it from the site http://extractor.iit.nrc.ca/ (21 January 2000). n-STEIN is a commercial system available for demonstration (http://www.gespro.com). It only produced abstracts at 5%, 10% or 15%.
In Figures 7.6 and 7.7, we present abstracts produced by **Extractor** and **n-STEIN** used in this evaluation.

![The Preci-Check flexible measuring system](image)

This paper explains the internal design of the Preci-Check measuring system and will discuss the collaboration of hardware and software. Human interface for application operation

The Preci-Check system required that an operator could train the robot, select programs or start and stop production as desired. The interface needed to allow for manual and automatic operation of the system with simple, user friendly software.

**Fig. 7.6: n-STEIN abstract for the Document “The Preci-Check flexible measuring system”, Industrial Robot, Vol 26 Issue 2, 1999**

![This paper explains the internal design of the Preci-Check measuring system](image)

This paper explains the internal design of the Preci-Check measuring system and will discuss the collaboration of hardware and software. The Preci-Check measuring system is a (6) axis, articulating robot arm on a longitudinal slide unit.

**Fig. 7.7: Extractor abstract for the Document “The Preci-Check flexible measuring system”, Industrial Robot, Vol 26 Issue 2, 1999**

**Materials, Subjects and Evaluation Procedure**

We used 15 source document from the journal Industrial Robots (the twelve source documents from the previous evaluation and three new documents). We produced 15 abstracts using our method and computed the compression ratio in number of words, then we produced abstracts by **n-STEIN** using a compression ratio as close as ours as possible. In most of the cases the abstracts produced by **n-STEIN** were more verbose than the abstracts produced with **SumUM**. We produced abstracts by **Extractor** by specifying the number of sentences to select. We produced forms following the same procedure than in the previous evaluation (see page 155). This time the forms contained different number of abstracts to accommodate to the number of subjects. We had a total of 23 evaluators. Our evaluators were 13 subjects from the McGill Graduate School of Library & Information Studies and 10 subjects from the program on Information Studies at John Abbott College. The evaluation procedure was the same as in the previous evaluation. For each abstract, we computed the same measures than in the previous evaluation (see page 154). The results for individual articles and the average information are shown in Table 7.11.
7.4. THIRD EVALUATION : INFORMATIVENESS

<table>
<thead>
<tr>
<th>Article</th>
<th>n-STEIN</th>
<th>Indicative</th>
<th>Quality</th>
<th>SumUM</th>
<th>Indicative</th>
<th>Quality</th>
<th>Extractor</th>
<th>Indicative</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>yes</td>
<td>3.17</td>
<td>yes</td>
<td>3.07</td>
<td>yes</td>
<td>3.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>yes</td>
<td>3.03</td>
<td>yes</td>
<td>3.67</td>
<td>no</td>
<td>1.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>no</td>
<td>1.67</td>
<td>no</td>
<td>1.33</td>
<td>no</td>
<td>4.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>yes</td>
<td>3.13</td>
<td>yes</td>
<td>4.00</td>
<td>yes</td>
<td>4.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>yes</td>
<td>3.33</td>
<td>yes</td>
<td>3.13</td>
<td>yes</td>
<td>3.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>no</td>
<td>3.40</td>
<td>no</td>
<td>3.70</td>
<td>no</td>
<td>4.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>yes</td>
<td>3.67</td>
<td>yes</td>
<td>3.67</td>
<td>yes</td>
<td>4.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>yes</td>
<td>1.67</td>
<td>yes</td>
<td>2.50</td>
<td>yes</td>
<td>3.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>no</td>
<td>2.00</td>
<td>no</td>
<td>3.50</td>
<td>yes</td>
<td>2.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>yes</td>
<td>3.17</td>
<td>yes</td>
<td>3.50</td>
<td>yes</td>
<td>3.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>no</td>
<td>1.00</td>
<td>yes</td>
<td>4.00</td>
<td>no</td>
<td>3.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>no</td>
<td>2.17</td>
<td>yes</td>
<td>3.60</td>
<td>yes</td>
<td>3.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>yes</td>
<td>3.00</td>
<td>yes</td>
<td>2.67</td>
<td>yes</td>
<td>2.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>yes</td>
<td>3.83</td>
<td>yes</td>
<td>2.10</td>
<td>yes</td>
<td>3.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>yes</td>
<td>3.17</td>
<td>yes</td>
<td>2.50</td>
<td>yes</td>
<td>4.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Average: 67% 2.76 80% 3.13 73% 3.47

Tab. 7.11: Results of Human Judgment about Indicativeness and Text Quality. Data from the evaluation carried out at John Abbott College

Results

Content: In 80% of the cases, the abstracts produced by SumUM were correctly classified by the evaluators. Instead, the abstracts produced by Extractor were correctly classified in 73% of the cases and the abstracts produced by n-STEIN were correctly classified in 67% of the cases.

Quality: On the average, all the automatic systems produced acceptable abstracts. In this evaluation, the abstracts produced by Extractor were of better quality than the abstract produced by SumUM and n-STEIN.

7.4 Third Evaluation: Informativeness

In this experiment we compared the sentences that SumUM selected in order to produce the indicative-informative abstracts with the sentences that human judges choose. This method of evaluation has already been used in other summarization evaluations such as (Edmunson, 1969; Salton et al., 1997; Marcu, 1997a; Jing et al., 1998). According to Salton, the idea is that if we find a high overlap between the sentences selected by an automatic method and the sentences selected by a human, the method can be regarded as effective. Nevertheless, the method has been criticized because of the low ratio of agreement between
human subjects in this task (Jing et al., 1998).

7.4.1 Materials

Source Documents: We used ten technical articles from two different sources: 5 from the Rapid Prototyping Journal and 5 from the Internet Research Journal. The documents are available on the Emerald Electronic Library and were downloaded in both plain text and HTML formats. The documents contain an average of 39K characters (minimum 24K and maximum 54K), an average of 236 sentences (minimum 173 and maximum 292) and an average of 5120 words (minimum 3209 and maximum 6907).

Summarization Systems: We considered three automatic system in this evaluation: SumUM, Microsoft’97 Summarizer, and Extractor.

Abstracts: We produced three abstracts for each document. First we produced abstract using SumUM (using as input the plain text). We counted the number of sentences selected by SumUM in order to produce the indicative-informative abstract (i.e., \( N_i = | \text{indicative content} \cup \text{informative data base} | \)). Then we produced other two automatic abstracts, one using Microsoft’97 Summarizer and other using Extractor. We specified to the system to select the same number of sentences as SumUM selected. This time we used the HTML formats, the abstracts, and lists of keywords provided with the journal were deleted from the source documents. In the case of Extractor, we considered as abstract only the list of sentences selected by the system discarding the list of keywords it produces. In this experiment we had a total of 30 automatic abstracts which can be seen as sets of sentences. To sum up, we have the following abstracts represented as sets of sentences:

- \( \text{Abstract(Extractor, } i \text{)} \) is the set of sentences selected by Extractor for document \( i \);
- \( \text{Abstract(Microsoft, } i \text{)} \) is the set of sentences selected by Microsoft’97 Summarizer for document \( i \); and
- \( \text{Abstract(SumUM, } i \text{)} \) is the set of sentences selected by SumUM for document \( i \).

Subjects: We had a total of 9 informants from the Laboratoire RALI de l’Université de Montréal, all of them researchers in Computational Linguistics. We relied on them to obtain an assessment about important sentences in the source documents.

Informative Ideal Abstracts: Each informant read two articles\(^7\) which were printed without abstracts and lists of keywords. The informant choose a number of important sentences from each article (until a maximum of \( N_i \), the number of sentences chosen by the summarization methods). Each article was read by two different informants, we thus had two sets of sentences for each article. We call these sets \( S_{i,j} (1 <= i <= 10 \land j = 1, 2) \). Most of the informants found the task quite complex.

\(^7\)One of the informants read four articles.
7.4.2 Procedure

We compared the sentences produced by each method with the sentences selected by the informants computing recall, precision and F-score (here Method is SumUM, Microsoft or Extractor):

- For the ideal abstract $S_{i,j}$, $\text{Recall}(\text{Abstract}(\text{Method},i),j)$ is the ratio of the number of relevant sentences identified by the automatic method Method (those appearing on both the ideal abstract $S_{i,j}$ and on the automatic abstract $\text{Abstract}(\text{Method},i)$) over the number of the relevant sentences identified by the human assessor $j$. This number is obtained using the following formula:

$$\text{Recall}(\text{Abstract}(\text{Method},i),j) = \frac{|S_{i,j} \cap \text{Abstract}(\text{Method},i)|}{|S_{i,j}|}$$

- For the ideal abstract $S_{i,j}$, $\text{Precision}(\text{Abstract}(\text{Method},i),j)$ is the ratio of relevant sentences identified by the automatic method Method (those appearing on both the ideal abstract $S_{i,j}$ and on the automatic abstract $\text{Abstract}(\text{Method},i)$) over the number of sentences identified by the automatic method. This number is obtained using the following formula:

$$\text{Precision}(\text{Abstract}(\text{Method},i),j) = \frac{|S_{i,j} \cap \text{Abstract}(\text{Method},i)|}{|\text{Abstract}(\text{Method},i)|}$$

- For the ideal abstract $S_{i,j}$, $\text{Fscore}(\text{Abstract}(\text{Method},i),j)$ is a composite score that combines the precision and recall measures. This number is obtained using the following formula:

$$\text{Fscore}(\text{Abstract}(\text{Method},i),j) = \frac{2 \cdot \text{Recall}(\text{Abstract}(\text{Method},i),j) \cdot \text{Precision}(\text{Abstract}(\text{Method},i),j)}{\text{Recall}(\text{Abstract}(\text{Method},i),j) + \text{Precision}(\text{Abstract}(\text{Method},i),j)}$$

The figures obtained are presented in Table 7.12: each row contains the information about an ideal abstract compared with each of the automatic methods.

In order to obtain a clear picture, we borrowed the scoring methodology proposed by Salton et al. (1997), additionally considering the following situations:

- union scenario: for each document we considered the union of the sentences selected by the two informants ($S_{i,1} \cup S_{i,2}$) and computed recall, precision, and F-score for each method;

- intersection scenario: for each document we considered the intersection of the sentences selected by the two informants ($S_{i,1} \cap S_{i,2}$) and computed recall, precision and, F-score for each method;

- optimistic scenario: for each document and method we considered the case in which the method performed the best (higher F-score) and computed recall, precision, and F-score. For example for the first document (see Table 7.12) we considered that SumUM and Microsoft’97 Summarizer performed better considering $S_{1,1}$; while Extractor performed better considering $S_{1,2}$;
<table>
<thead>
<tr>
<th>Document</th>
<th>SumUM</th>
<th>Microsoft</th>
<th>Extractor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R  P  F</td>
<td>R  P  F</td>
<td>R  P  F</td>
</tr>
<tr>
<td>S₁,₁</td>
<td>.08 .15 .11</td>
<td>.05 .10 .07</td>
<td>.03 .06 .04</td>
</tr>
<tr>
<td>S₂,₁</td>
<td>.03 .05 .04</td>
<td>.03 .05 .04</td>
<td>.16 .31 .21</td>
</tr>
<tr>
<td>S₂,₂</td>
<td>.24 .23 .24</td>
<td>.24 .18 .20</td>
<td>.10 .20 .13</td>
</tr>
<tr>
<td>S₃,₁</td>
<td>.25 .18 .21</td>
<td>.38 .21 .27</td>
<td>.09 .15 .12</td>
</tr>
<tr>
<td>S₃,₂</td>
<td>.24 .23 .23</td>
<td>.05 .05 .05</td>
<td>.05 .06 .05</td>
</tr>
<tr>
<td>S₄,₁</td>
<td>.38 .38 .38</td>
<td>.10 .09 .09</td>
<td>.12 .24 .16</td>
</tr>
<tr>
<td>S₄,₂</td>
<td>.25 .25 .25</td>
<td>.17 .16 .17</td>
<td>.11 .24 .15</td>
</tr>
<tr>
<td>S₅,₁</td>
<td>.20 .21 .21</td>
<td>.29 .30 .29</td>
<td>.06 .09 .07</td>
</tr>
<tr>
<td>S₅,₂</td>
<td>.11 .09 .10</td>
<td>.30 .24 .27</td>
<td>.04 .04 .05</td>
</tr>
<tr>
<td>S₆,₁</td>
<td>.54 .25 .35</td>
<td>.08 .04 .05</td>
<td>.19 .21 .20</td>
</tr>
<tr>
<td>S₆,₂</td>
<td>.22 .20 .21</td>
<td>.18 .17 .17</td>
<td>.20 .42 .27</td>
</tr>
<tr>
<td>S₇,₁</td>
<td>.23 .09 .13</td>
<td>.08 .03 .04</td>
<td>.00 .00 . -</td>
</tr>
<tr>
<td>S₇,₂</td>
<td>.25 .18 .21</td>
<td>.08 .05 .06</td>
<td>.29 .35 .32</td>
</tr>
<tr>
<td>S₈,₁</td>
<td>.10 .08 .09</td>
<td>.03 .03 .04</td>
<td>.06 .08 .07</td>
</tr>
<tr>
<td>S₈,₂</td>
<td>.11 .11 .11</td>
<td>.09 .08 .08</td>
<td>.11 .17 .14</td>
</tr>
<tr>
<td>S₉,₁</td>
<td>.25 .22 .24</td>
<td>.13 .09 .10</td>
<td>.38 .45 .41</td>
</tr>
<tr>
<td>S₉,₂</td>
<td>.38 .19 .25</td>
<td>.00 .00 .00</td>
<td>.08 .05 .06</td>
</tr>
<tr>
<td>S₁₀,₁</td>
<td>.19 .19 .19</td>
<td>.13 .05 .07</td>
<td>.13 .17 .14</td>
</tr>
<tr>
<td>S₁₀,₂</td>
<td>.25 .25 .25</td>
<td>.13 .05 .07</td>
<td>.13 .17 .14</td>
</tr>
<tr>
<td>Average</td>
<td>.23 .20 .21</td>
<td>.14 .11 .12</td>
<td>.12 .18 .14</td>
</tr>
</tbody>
</table>

Tab. 7.12: Comparison between sentences selected by human informants and sentences selected by three automatic summarization methods: general scenario. The columns contain the information about Recall, Precision and F-score.

- pessimistic scenario: for each document and method we considered the case in which the method performed the worst (lower F-score) and computed recall, precision, and F-score. For example, for the first document (see Table 7.12) we considered that **SumUM** and **Microsoft‘97 Summarizer** performed worse considering **S₁,₂**; while **Extractor** performed worse considering **S₁,₁**.

For each scenario we computed the average information which is presented in Table 7.13.

### 7.4.3 Results of Summarization Systems

For the scenario in which we consider the 20 human abstracts (i.e., two different abstracts for each of the ten documents), **SumUM** obtained better F-score in 60% of the cases, **Extractor** in 25% of the cases, and **Microsoft‘97 Summarizer** in 15% of the cases. If we assume that the sentences selected by the human informants represent the most important or interesting information of the documents, then we can conclude that in most of the cases **SumUM** performed better than the two other summarization technologies. Even if these results are not exceptional in individual cases, **SumUM** performed better than the other summarization methods on the average. The average F-score over the 20 abstracts is 21%
### 7.4. Third Evaluation: Informativeness

<table>
<thead>
<tr>
<th></th>
<th>SumUM</th>
<th></th>
<th></th>
<th>Microsoft</th>
<th></th>
<th></th>
<th>Extractor</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>Union</td>
<td>.21</td>
<td>.31</td>
<td>.25</td>
<td>.16</td>
<td>.19</td>
<td>.17</td>
<td>.11</td>
<td>.26</td>
<td>.15</td>
</tr>
<tr>
<td>Intersection</td>
<td>.28</td>
<td>.09</td>
<td>.14</td>
<td>.13</td>
<td>.04</td>
<td>.06</td>
<td>.08</td>
<td>.04</td>
<td>.06</td>
</tr>
<tr>
<td>Pessimistic</td>
<td>.19</td>
<td>.17</td>
<td>.18</td>
<td>.11</td>
<td>.08</td>
<td>.09</td>
<td>.08</td>
<td>.11</td>
<td>.09</td>
</tr>
</tbody>
</table>

Tab. 7.13: Comparison between sentences selected by human informants and sentences selected by three automatic summarization methods: union, intersection, optimistic and pessimistic scenarios. The columns contains the information about average Recall, Precision and F-score representing a gain of 50% over Extractor and a gain of 75% over Microsoft’97 Summarizer.

In the different proposed scenarios, SumUM performed better than the other two methods. For the union scenario, SumUM obtained a better F-score in 70% of the cases, Microsoft’97 Summarizer in 20% of the cases, and Extractor in 10% of the cases. For the intersection scenario, SumUM obtained a better F-score in 40% of the cases, Microsoft’97 Summarizer in 20% of the cases, and Extractor in 20% of the cases. There were no matches for two documents. For the optimistic scenario, SumUM and Extractor obtained a better F-score in 40% of the cases and Microsoft’97 Summarizer in 20% of the cases. In the pessimistic scenario, SumUM obtained a better performance in 90% of the cases, and Microsoft’97 Summarizer in 20% of the cases (both systems performed alike in one case). Extractor had no better F-score in this scenario. In the pessimistic scenario, Extractor obtained better performance in precision over the other two methods with a gain of 9% over SumUM and a gain of 79% over Microsoft’97 Summarizer.

Here, we have compared three different methods of producing abstracts that are domain independent. Nevertheless, while Microsoft’97 Summarizer and Extractor are truly genre independent, SumUM is genre dependent: it was designed for technical articles and takes advantage of this fact in order to produce the abstracts. We think that this is the reason for the better performance of SumUM in this evaluation. The results of this experiment are encouraging considering the limited capacities of the actual implementation. We expect to improve the results in future versions of SumUM.

#### 7.4.4 Comparison of Human Summaries

We have also compared the sentences selected by the two groups as follows: first, we considered the sentences selected by the first group as ideal abstract and we computed recall, precision and F-score for the second group, then we interchanged the first group with the second group and computed recall, precision and F-score for the first group. The results are presented in Table 7.14. The first column considers the first group as ideal abstract while the second column considers the second group as ideal. In 80% of the cases the human group
was better than the automatic abstracts. Even if the average overlap is only 37% (Salton et al. (1997) found it around 46% in their experiments), the abstracts produced by a human group have higher overlap with the ideal abstract than any of the automatic abstracts.

<table>
<thead>
<tr>
<th># Doc.</th>
<th>Group II</th>
<th></th>
<th>Group I</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>1</td>
<td>.27</td>
<td>.32</td>
<td>.29</td>
<td>.32</td>
<td>.27</td>
</tr>
<tr>
<td>2</td>
<td>.66</td>
<td>.51</td>
<td>.58</td>
<td>.51</td>
<td>.66</td>
</tr>
<tr>
<td>3</td>
<td>.43</td>
<td>.41</td>
<td>.42</td>
<td>.41</td>
<td>.43</td>
</tr>
<tr>
<td>4</td>
<td>.38</td>
<td>.38</td>
<td>.38</td>
<td>.38</td>
<td>.38</td>
</tr>
<tr>
<td>5</td>
<td>.31</td>
<td>.41</td>
<td>.35</td>
<td>.41</td>
<td>.31</td>
</tr>
<tr>
<td>6</td>
<td>.28</td>
<td>.54</td>
<td>.37</td>
<td>.54</td>
<td>.28</td>
</tr>
<tr>
<td>7</td>
<td>.15</td>
<td>.08</td>
<td>.11</td>
<td>.08</td>
<td>.15</td>
</tr>
<tr>
<td>8</td>
<td>.46</td>
<td>.52</td>
<td>.48</td>
<td>.52</td>
<td>.46</td>
</tr>
<tr>
<td>9</td>
<td>.31</td>
<td>.17</td>
<td>.22</td>
<td>.17</td>
<td>.31</td>
</tr>
<tr>
<td>10</td>
<td>.31</td>
<td>.31</td>
<td>.31</td>
<td>.31</td>
<td>.31</td>
</tr>
<tr>
<td>Average</td>
<td>.37</td>
<td>.38</td>
<td>.37</td>
<td>.38</td>
<td>.37</td>
</tr>
</tbody>
</table>

Tab. 7.14: Comparison between sentences selected by the first group and sentences selected by the second group. Each row represents a document. The last row is the average information. The columns contain the information about Recall, Precision and F-score.

### 7.5 Discussion

In our first evaluation, we measure how many of the “topics” of the source document are covered by the automatic abstracts by comparing the terms of the automatic abstracts with the terms of an ideal abstract. In order to cope with the lack of evaluation resources for the technical domain, we relied on technical documents found on the Web and we used the abstract provided with the source document as an ideal abstract. Sometimes those abstracts fail to indicate the essential content of the document. In fact, we had to exclude some articles from our test set because the terms appearing in the provided abstracts were not found in the technical document and as a result all the three methods failed to indicate the “topics” of the source text. But, the fact that a term appearing in the provided abstract didn’t appear in the source document doesn’t mean that it is not a topic. That term could be obtained using a deductive process.

Unfortunately in this experiment, we have chosen to extract terms from the abstracts using the same term extraction algorithm developed for SumUM. We have verified that the use of that algorithm gave an unexpected advantage to SumUM. In an evaluation using nouns instead of terms as representation of the abstracts, we obtained the same ranking for the systems but the differences are smaller than in our previous evaluation using terms. This last experience showed us the need for evaluation procedures not biased towards any particular methodology.
7.5. DISCUSSION

Our second evaluation measured how helpful automatic abstracts are in a classification task: evaluators were asked to classify abstract in categories represented by lists of keywords (descriptors). We were motivated by the following fact, suppose that we have two articles in the field of robotics, one talking about “cooking robots” and other talking about “janitor robots”, then those topics will probably appear in the descriptors for those articles. Suppose that we want to know (without looking to the articles) which talks about which subject by examination of their abstracts. If the abstract allows readers to discriminate between the two articles (because the topic is explicitly stated or because it allows some deduction of the topic), we could say that it was helpful in the classification task. All the automatic methods performed similarly, though, we believe that documents and descriptors of narrowed domains are needed in order to correctly assess the effectiveness of each summarization method. Unfortunately, the construction of such resources goes beyond our present research and will be addressed in future work. We took advantage of the experiments to ask readers to evaluate the quality of the text based on a number of criteria generally accepted in abstract writing (the criteria were not enforced). The results indicate that the abstracts produced by SumUM are acceptable when compared with a sentence extraction system. Nevertheless, this is the case when the compression ratio is high (more than 90%) because the extraction based system chooses indicative elements such as titles and captioning that make the text difficult to read, while SumUM always presents complete sentences to the reader. In the third run of this evaluation, Extractor produced abstracts considered slightly better than the abstracts by SumUM. The low quality score and the few comments of the evaluators about the abstracts produced by SumUM indicate that much work is needed in the regeneration step in order to produce truly good quality abstracts. Unfortunately, we could not compare the abstracts by SumUM with other automatic abstracting systems simply because they are not available. Minel et al. (1997) have proposed two methods of evaluation addressing the content of the abstract and its quality. For content evaluation, they asked human judges to classify summaries in broad categories (politics, science and technique, etc.) and also verify if the key ideas of source documents are appropriately expressed in the summaries. For text quality, they asked human judges to identify problems such as dangling anaphora and broken textual segments and also to make subjective judgments about readability.

Our evaluation mirrors the TIPSTER SUMMAC categorization task (Firmin and Chrzansowski, 1999; Mani et al., 1998) in which given a generic summary (or a full document), the human subject chooses a single category (out of five categories) to which the document is relevant. The evaluation seeks to determine whether the summary is effective in capturing whatever information in the document is needed to correctly categorize the document. In that evaluation, 10 TREC topics and 100 documents per topic were used. In that evaluation 16 systems participated. The results indicate that there are not significant differences among the systems for the categorization task and that the performance using the full document is not much better. For text quality, the SUMMAC evaluation addressed subjective aspects such as the length of the summary (i.e., too short, too long, etc.), its intelligibility (i.e., poor, OK, etc.) and its usefulness (i.e., low, high, etc.).

Finally, our third evaluation addressed the issue of informativeness. We have compared
the sentences that SumUM selected for the indicative-informative abstract with sentences considered important by human informants. We found that SumUM performed better than two other summarization technologies in the different scenarios proposed by Salton et al. (1997). Nevertheless, SumUM performed worst than the human abstracts, so an additional effort is needed in order to improve the informative content of our abstracts. The evaluation considered only ten documents, this could be seen as an small evaluation, though, it is comparable in volume with other evaluations recently done in text summarization (Jing et al., 1998) using more texts but of smaller size.

7.6 Summary

In this chapter, we presented three evaluations methodologies of the abstracts produced by SumUM. In the three evaluations, we have compared different summarization technologies. In the evaluation of indicative content, no differences were observed among different summarization systems in a classification task. Regarding the content of the indicative-informative abstracts, we found that on the average SumUM selected sentences that are more important to the readers than sentences selected by other summarization technologies. Regarding text acceptability, the indicative abstracts are acceptable texts given a set of criteria usually used in abstract writing, though, the low score obtained by SumUM, calls for the improvement of the regeneration step. How to improve the text will be subject of our future work.
Chapitre 8

Conclusion et Perspectives

Comment faire pour qu’un ordinateur calcule le contenu essentiel d’un document et l’exprime sous la forme d’un nouveau texte cohérent et cohérent ? Telle est la problématique du résumé automatique abordée dans cette thèse. Produire des résumés est une tâche fort difficile, car elle nécessite des connaissances linguistiques et du monde lesquelles ne sont pas faciles à incorporer dans des systèmes automatiques.

Afin de comprendre comment les résumés automatiques sont produits, nous avons dans un premier temps abordé l’étude des méthodes existantes pour résoudre ce problème. Nous avons constaté l’intérêt des approches statistiques qui sont relativement faciles à implanter et indépendantes du domaine ; les résumés obtenus sont dans la plupart des cas des extraits de texte. Nous avons aussi constaté que les approches symboliques qui s’appuient sur des connaissances linguistiques et du monde pour produire de vrais textes ne peuvent être appliquées que dans des domaines restreints et avec beaucoup d’effort car les connaissances doivent être généralement acquises manuellement.

Dans notre recherche, nous avons essayé de limiter la complexité de la tâche en nous concentrant sur un seul type de résumé : le résumé indicatif-informatif du texte technique et scientifique. Néanmoins, nous n’avons pas limité le domaine textuel parce que le texte technique, peu importe son domaine, fait toujours référence à des informations générales du monde conceptuel de la recherche. Comme nous voulions une base valable pour le développement d’un système informatique de génération de résumés, nous avons abordé les questions du calcul du contenu et de sa reformulation en étudiant un corpus de résumés rédigés par des rédacteurs professionnels et leurs documents sources. Ce choix qui a influencé notre démarche ultérieure est basé sur le fait que les résumés professionnels sont des textes bien structurés et qui préservent les mots de l’auteur. Cette étude nous a mené vers la définition d’un modèle linguistique et conceptuel pour le résumé du texte technique et scientifique. Nous avons montré quelles sont les parties structurales du document source le plus souvent utilisées pour repérer le contenu essentiel ; nous avons aussi étudié quel est le statut de ce contenu et comment il est reformulé sous la forme d’un nouveau texte. Nous avons ainsi étudié la relation entre le texte source et le résumé dans le cas du texte technique et scientifique.
In faut noter que dans le domaine du résumé automatique l’étude de la relation entre texte source et résumé a été généralement négligée. En conséquence, notre étude de corpus et le modèle linguistique et conceptuel qui en découlent constituent une contribution dans cette direction. Les concepts, relations, types d’information, marqueurs linguistiques et modèles que nous avons trouvés peuvent être utilisés comme base pour sélectionner l’information pertinente dans un document technique. Ici, il est important de laisser clair que la question de la couverture du modèle proposé n’a pas été traitée dans cette thèse et constitue un sujet de recherche prometteur. Cependant, il est clair que la liste des expressions linguistiques et des modèles linguistiques n’est pas complète. Il n’est pas difficile d’imaginer des exemples de types d’informations présentes dans le modèle mais pour lesquelles nous n’avons pas de marqueur linguistique : par exemple, une phrase comme “This approach reduces development costs” parle, certainement, d’un avantage ; par contre une phrase comme “This new treatment reduces life expectancy” indique, plutôt, une désavantage. Les phrases sont semblables en structure, la même construction linguistique “X reduces Y” est utilisée, pourtant seule la sémantique des arguments du verbe “reduce” aide choisir l’interprétation valide.

La tâche de trouver des expressions linguistiques est laborieuse et nécessite la disponibilité d’un volume considérable de résumés électroniques. De nos jours, on peut obtenir ces ressources, par exemple dans le site que nous avons exploré pour nos expériences (EMERALD, 2000), dans des sites dédiés aux publications scientifiques (SCIENCE DIRECT, 2000) et même dans des services de diffusion des résumés en ligne (INSPEC, 2000). L’acquisition de connaissances linguistiques à partir de ces ressources peut être atteinte avec des méthodes semi-automatiques pour réduire le temps de développement tel que proposé par Teufel (1998).

Bien que certains concepts, relations et types d’information soient distinctifs du texte technique, les types d’information utilisés pour le résumé informatif sont généraux et peuvent être utilisés dans d’autres genres textuels. Néanmoins, ces informations restent insuffisantes pour la production de résumés informatifs pour des lecteurs des domaines spécifiques ; par exemple, en Micro-biologie, un lecteur pourrait être intéressé aux informations sur les matériels utilisés dans les expériences et donc il aura besoin de trouver des phrases comme “Bacterial cells were collected by centrifugation, washed three times with pyrogen-free water, and lyophilized.” qui ne peuvent pas être repérés par notre modèle. Ceci reste un problème à explorer. À notre connaissance, notre analyse des transformations appliquées au texte source afin de produire un résumé n’a jamais été étudié auparavant. Pourtant, notre travail et d’autres travaux récents montrent qu’il s’agit d’une avenue qui doit être explorée (Jing and McKeown, 2000; Knight and Marcu, 2000). Les limitations de notre étude sont toutefois nombreuses car nous n’avons exploré que des transformations au niveau de la phrase. Au niveau du texte, nous avons proposé un ordre de présentation de l’information qui est typique dans le résumé technique, pourtant une étude se fait nécessaire afin d’identifier les structures discursives valides et le choix de la plus approprié pour le résumé technique. Ceci sera l’un des aspects à explorer dans notre travail futur tout en utilisant notre corpus de résumés.

Nous avons défini l’Analyse Sélective, une méthodologie pour la production automatique
de résumés indépendamment de son implantation, en considérant pourtant les capacités actuelles du traitement du langage naturel. La méthode propose dans un premier temps la production d’un résumé indicatif qui permet au lecteur de connaître les principaux sujets traités dans le document source. Dans un deuxième temps, le lecteur peut obtenir sur demande des informations supplémentaires à propos des thèmes principaux. Nous considérons que les thèmes principaux abordés dans un document peuvent être calculés en considérant un nombre limité de concepts et de relations et en s’appuyant sur la structure du document. Nous explorons ainsi l’approche des résumés dynamiques dont le contenu change selon les intérêts de l’utilisateur. Néanmoins, nous n’avons considéré qu’un nombre limité de types d’information auxquels l’utilisateur pourrait s’intéresser, la question d’adapter le résumé aux besoins particuliers d’un utilisateur n’a pas été abordée dans cette thèse. Pour représenter les types d’information auxquels nous nous intéressons, nous avons défini des templates que nous utilisons pour les tâches de sélection de contenu et régénération. Nous avons associé aux templates des schémas de régénération de textes qui rendent possible l’implantation des phénomènes observés lors de notre analyse de corpus. D’autres structures de représentation plus riches peuvent être nécessaires pour améliorer les résultats, néanmoins, pour la recherche décrite dans cette thèse ils se sont avérés appropriés.

Dans l’Analyse Sélective, l’information pour le résumé dépend de la structure de titres du document et des modèles spécifiques que les phrases doivent respecter pour être sélectionnées. Bien que ces dépendances aident à réduire le “bruit” dans le calcul du contenu et à rendre possible la régénération des phrases, elles contribuent aussi au “silence” : beaucoup d’information est perdue à cause de ses dépendances, il faudrait explorer d’autres techniques de sélection de contenu, tel que la combinaison de l’évidence indicative et informative.

Afin de montrer la viabilité de notre approche nous avons développé SumUM un système informatique qui implante les fonctionnalités de l’Analyse Sélective tout en utilisant des méthodes robustes de traitement du langage naturel. Notre implantation se base sur l’utilisation des programmes d’étiquetage lexical, l’application des automates à états finis pour reconnaître des constructions syntaxiques simples, la classification des phrases, l’extraction d’informations, l’instanciation de patrons et la régénération de textes. Ces techniques nous ont aidé à implanter quelques phénomènes observés dans le corpus tels que la sélection de fragments de phrases, la fusion de l’information et la reformulation, aspects généralement ignorés dans les systèmes de génération des résumés. Nous avons implanté SumUM avec des algorithmes simples en Prolog. Les différentes composantes du système ont été évaluées de manière subjective pendant les tests que nous avons effectués, pourtant une évaluation formelle serait nécessaire afin de mettre en évidence les points faibles et les potentiels de SumUM. Parmi les faiblesses de cette implantation se trouvent l’utilisation de modèles parfois ambigus, le manque de connaissances lexicales (thesaurus), la simplicité du processus de génération et le fait que l’acquisition des connaissances (lexique et modèles) se fasse de manière manuelle. Dans nos travaux futurs, nous désirons explorer l’utilisation de méthodes d’acquisition automatique de connaissances et de méthodes de classification statistique afin de surmonter les faiblesses actuelles. Nous voulions transformer quelques fonctionnalités de SumUM, tels que la sélection de phrases pour le résumé indicatif et l’expansion informative,
dans un système “réel” qui puisse être utilisée sur le Web pour produire des résumés dans d’autres domaines que techniques et dans d’autres langues que l’anglais.

Nous avons évalué notre modèle théorique en utilisant les résumés produits par **SUMUM** et en les comparant avec des résumés humains et des résumés produits par d’autres techniques et systèmes actuels de production de résumés par extraction de phrases. Trois aspects des résumés ont été évalués : le contenu indicatif, le contenu informatif et la qualité du texte. Du point de vue du contenu indicatif et informatif nous avons vérifié que **SUMUM** produit des résumés dont le contenu est en moyenne plus pertinent que le contenu produit par d’autres systèmes automatiques. Toutefois, dans l’évaluation du contenu indicatif dans une tâche de classification nous n’avons pas observé des différences parmi les différentes méthodes de production de résumés, et donc, des évaluations avec des catégories plus restrictives se font nécessaires. Du point de vue de la qualité du texte, nous avons constaté que les résumés sont d’une qualité acceptable et parfois comparable aux résumés humains. Néanmoins, les quelques critiques obtenues lors de l’évaluation avec des juges et les notes obtenues indiquent des avenues pour l’amélioration du prototype à explorer. Toutefois, nous croyons que le modèle théorique est valable. Quoique notre évaluation n’ait porté que sur un nombre relativement restreint de documents (57 documents), 36 juges ont participé à nos évaluations. Même si nous avons exploré l’utilisation de ressources électroniques disponibles sur le Web pour notre évaluation, la création de ressources et de critères objectifs d’évaluation de résumés de science et technologie reste un thème à explorer. Même si les programmes d’évaluation tel que TIPSTER SUMMAC de l’agence DARPA (Defense Advance Research Projects Agency) se concentrent sur des textes utilisés par des analystes de la défense, et donc textes non scientifiques, nous espérons que dans l’avenir des ressources pour l’évaluation des résumés techniques seront disponibles.

Il est important de mentionner qu’un thème qui commence à être exploré en linguistique informatique est celui de la production des résumés de plusieurs articles (multi-document) qui traitent du même sujet et même dans plusieurs langues (multi-langue). Le programme TIDES (Translingual Information Detection, Extraction and Summarization) de la DARPA a été récemment lancé pour faire avancer la recherche dans cette direction (TIDES, 2000). Mais, il ne faut pas être naïf, le fait qu’il y ait de groupes qui se lancent sur le problème du résumé multi-langue et sur le résumé multi-document ne veut pas dire que le problème de produire un seul résumé pour un seul document dans une seule langue ait été déjà résolu.

Finalement et pour conclure, il semble que l’avenue à explorer dans le domaine de la génération de résumés est celle de la combinaison de différentes méthodes, l’incorporation d’outils robustes de traitement du langage naturel, l’incorporation de techniques de régénération pour améliorer la qualité du texte et l’acquisition automatique des connaissances afin d’obtenir des algorithmes qui peuvent s’adapter facilement à des nouveaux domaines.

Dans cette recherche nous avons parcouru le long et ardu chemin qui mène du texte source au résumé : nous avons étudié la question du contenu du résumé et de sa forme, nous
avons conçu et implanté une nouvelle méthode pour le résumé automatique et nous l'avons validée. Nous avons constaté qu'on est encore loin de produire des résumés comparables aux résumés produits par un humain. Plusieurs questions toutefois n’ont été que superficiellement explorées : la complétude de notre modèle théorique, les possibilités d’extensions du modèle vers d’autres types de texte et l’implantation d’un système qui puisse être utilisé dans notre travail pour faciliter l’accès à l’information. Ces questions seront l’objet de nos travaux futurs afin de produire de bons résumés entre autres pour la petite Margie.
Bibliographie


Annexe A

Journaux source utilisés pour le corpus

AI Communications
AI Magazine
American Libraries
Annals of Library Science & Documentation
Archives and Museum Informatics
Artificial Intelligence
Aslib Proceedings
Australian Library Journal
Bottom Line
Byte
Collection Management
College & Research Libraries
Computer Communications
Computer Networks and ISDN Systems
Computers in Libraries
Digital Publishing Technologies
Document Delivery & Information Supply
Electronic Library
Electronic Publishing
IATUL Proceedings
IEEE Expert
Information Outlook
International Journal of Human-Computer Studies
Interacting with Computers
Journal of End User Computing
Journal of Government Information
Journal of Information Science
Journal of Interlibrary Loan
Journal of Library Administration
Knowledge Based Systems
Library Administration & Management
Library Association Record
Library Hi Tech News
Library Journal
Libri
Microform & Imaging Review
New Review of Hypermidia and Multimidia
New Review of Information & Library Research
New Review of Information Science & Library Research
OCLC Newsletter
Scandinavian Public Library
Telematics and Informatics
Vine
Annexe B

Expressions indicatives identifiées
dans le corpus

we present in this paper section x we say that this paper is concerned with it is shown that the problem of In this paper we present we will first the results are reported of the possible reasons for ... are explored experiment x studied we report work the hypotheses under test evaluates we address the issue of we highlight it is argued I refer to seems to it will show was built the aim of this article is to compare in this paper we have presented we analyse the idea in this paper in Section x it is shown is investigated is discussed is considered has been proven to be very successful the basic idea is that we will then describe were compared the purpose of the experiments described here experiment x addressed issues are compared does not explain results shown we’ve conducted are reviewed the purpose of this paper have the potential the work described in this paper to this end it is remarkable this paper considers
<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>an issue</td>
<td>the fact is that</td>
</tr>
<tr>
<td>it is impossible</td>
<td>it is necessary</td>
</tr>
<tr>
<td>is heavily impacting</td>
<td>the motivation for our work is</td>
</tr>
<tr>
<td>a research project</td>
<td>is investigating</td>
</tr>
<tr>
<td>the aim is</td>
<td>currently work is focused</td>
</tr>
<tr>
<td>this has resulted</td>
<td>the advantage that</td>
</tr>
<tr>
<td>we study</td>
<td>we present</td>
</tr>
<tr>
<td>are studied</td>
<td>are compared</td>
</tr>
<tr>
<td>this article surveys</td>
<td>these problems</td>
</tr>
<tr>
<td>after a brief introduction</td>
<td>I outline</td>
</tr>
<tr>
<td>solutions</td>
<td>are being developed</td>
</tr>
<tr>
<td>describe that is aimed</td>
<td>its charter</td>
</tr>
<tr>
<td>finally we argue</td>
<td>I define</td>
</tr>
<tr>
<td>it has been conducting research and development</td>
<td>has become a powerful</td>
</tr>
<tr>
<td>I will discuss</td>
<td>in this presentation</td>
</tr>
<tr>
<td>we have focused on</td>
<td>suggests that</td>
</tr>
<tr>
<td>discussed how</td>
<td>aims</td>
</tr>
<tr>
<td>the study</td>
<td>suggest</td>
</tr>
<tr>
<td>this paper will consider</td>
<td>to this end</td>
</tr>
<tr>
<td>this paper describes the project</td>
<td>in this paper we present the results</td>
</tr>
<tr>
<td>our experiment</td>
<td>we found</td>
</tr>
<tr>
<td>the results indicate that</td>
<td>a papers based on</td>
</tr>
<tr>
<td>this paper presents</td>
<td>it shows</td>
</tr>
<tr>
<td>this paper combines</td>
<td>the work described in this paper</td>
</tr>
<tr>
<td>addresses issues of</td>
<td>in this paper is demonstrated</td>
</tr>
<tr>
<td>this paper describes a project</td>
<td>an analysis was undertaken</td>
</tr>
<tr>
<td>the sampling procedure is described</td>
<td>attention is focused on</td>
</tr>
</tbody>
</table>
the analytical tool which is used for this purpose is based on
the paper concludes
this brief paper attempts an overview
the ... project
the principal distinguishing features of
for this reason
the aim of the ... project is to
the project involves
the article describes objectives of
a key focus of ... was
the aim of ... are to
this paper ... but focusses on
I take a look
this paper discusses
the results of the survey are
those contributions represent
...began her presentation
...suggest using the
research in
this article highlights results
the first focus in on...next
this article describes
special emphasis on
...evaluation
the project demonstrates
it is essential that
the author have developed
it is one of a set of four projects
the overall objective of
... is one of four ... projects supported under
the ... project
the article describes the aims ...
... is a ... project founded
to develop
special attention is paid to
this initial description
a general survey was conducted
according to the survey
...presented her views on
...suggests the use of
another benefit of
this project is aimed at
...are examined in this column
this column will close with
this paper presents
quality tests were completed
the paper will touch on
the objective is both
the project...
the paper also focus on
...believes that
...are a major underlying themes
...will be given
...were conducted
...was examined
what follows are sketches of
the journal’s scope includes
we briefly overview some interesting
in this work we study
we classified
our conclusion was that
details of...are provided
this article discusses
other examples
overcoming many of
...described here
...for solving
this analysis summarizes
I give particular attention to
...to develop
after ...we describe

...are studied and applied
...developed and studied
...investigated
...outperform
...is emerging

the aim of
the subject of this paper is
the primary objective of this paper is
a number of suggestions are
in the next pages
it points to
this report describes case studies of
...selected for the study
interviews were carried out with
...enhanced
...considers
then we update
we look
we studied
...provides

the authors
the reasons
the proposed model
this article presents
this article describes
finally, I investigate
The ... project
this article reviews
we discuss the development and implementa-
tion
...model
...optimizes
this article assesses
...supports experimentation
## Annexe C

### Concepts

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Explanation &amp; Example</th>
<th>Lexical Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>author</td>
<td>The authors of the article.</td>
<td>We, I, author,...</td>
</tr>
<tr>
<td></td>
<td>“I refer to ...”</td>
<td></td>
</tr>
<tr>
<td>institutions</td>
<td>Institutions</td>
<td>University, Université,...</td>
</tr>
<tr>
<td></td>
<td>“Department of Mechanical and...”</td>
<td></td>
</tr>
<tr>
<td>affiliation</td>
<td>The affiliation of the authors.</td>
<td>Prop.Noun, Institution,...</td>
</tr>
<tr>
<td>researcher</td>
<td>Other researchers</td>
<td>Proper Noun...</td>
</tr>
<tr>
<td></td>
<td>“Cannon shown...”</td>
<td></td>
</tr>
<tr>
<td>author related</td>
<td>Authors’ related entity.</td>
<td>our, my, ...</td>
</tr>
<tr>
<td></td>
<td>“The core of our system is comprised of...”</td>
<td></td>
</tr>
<tr>
<td>research group</td>
<td>A research group.</td>
<td>group, ...</td>
</tr>
<tr>
<td></td>
<td>“… is a staff programmer in the Experimental Systems group at the IBM Corporation.”</td>
<td></td>
</tr>
<tr>
<td>project</td>
<td>A (research) project.</td>
<td>project, ...</td>
</tr>
<tr>
<td></td>
<td>“A research project currently in progress at...”</td>
<td></td>
</tr>
<tr>
<td>research paper</td>
<td>The technical article</td>
<td>article, here, paper,...</td>
</tr>
<tr>
<td></td>
<td>“In this article...”</td>
<td></td>
</tr>
<tr>
<td>others’ paper</td>
<td>The article of other researchers.</td>
<td>article, ...</td>
</tr>
<tr>
<td></td>
<td>“In their article...”</td>
<td></td>
</tr>
<tr>
<td>study</td>
<td>The object of study.</td>
<td>study, ...</td>
</tr>
<tr>
<td></td>
<td>“An empirical study of randomly generated binary ...”</td>
<td></td>
</tr>
<tr>
<td>research</td>
<td>The research work.</td>
<td>research, ...</td>
</tr>
<tr>
<td></td>
<td>“… a broad range of scientific research...”</td>
<td></td>
</tr>
<tr>
<td>work</td>
<td>The work of the author.</td>
<td>work, ...</td>
</tr>
<tr>
<td></td>
<td>“The work described in this paper addresses...”</td>
<td></td>
</tr>
</tbody>
</table>
| **problem** | The problem under consideration  
*“The lack of a library severely limits the impact of...”* | **difficulty, issue, problem, ...** |
|-------------|-------------------------------------------------------------------------------------------------|----------------------------------|
| **solution** | The solution to the problem  
*“In this paper, we describe the problem and propose a solution, ...”* | **solution, answer, ...** |
| **method** | The method used in the study  
*“One approach is to support indexing by the traditional method of assigning...”* | **equipment, methodology, ...** |
| **results** | The results obtained  
*“The results indicate that ...”* | **result, outcome, ...** |
| **experiment** | The experiment  
*“The experiments were done in order ...”* | **experiment, test, ...** |
| **need** | A necessity.  
*“...the need for an interface between ...”* | **need, necessity, ...** |
| **application** | Applicability.  
*“...the applicability of our approach ...”* | **use, employ, ...** |
| **hypothesis** | The hypothesis.  
*“That is, in other words, to get the most reasonable hypothesis.”* | **hypothesis, assumption, ...** |
| **question** | Research question.  
*“...to understand how services which are of value to end-users can be developed.”* | **research question, ...** |
| **future plans** | The future plans of the author.  
*“...some future work in the field.”* | **future plan, ...** |
| **reference** | Reference to previous work.  
*“...the Manhattan street network [4,5].”* | ** Proper Noun (Year), ...** |
| **acronym** | An acronym  
*“The World Wide Web (WWW)....”* | **Noun Group (Acronym), ...** |
| **acronym expansion** | The expansion of the acronym.  
*“The World Wide Web (WWW)....”* | **Noun Group (Acronym), ...** |
| **structural** | Structural element of the document such as a table or figure.  
*“In Figure 3 we show...”* | **figure, table, picture, ...** |
| **title** | A title from the document.  
*“2.3. Algorithm Implementation”* | ** Number Title, ...** |
| **captioning** | The captioning of a structural element.  
*“Fig. 2: Test Results.”* | ** Figure X Captioning, ...** |
| **quantity** | A quantity.  
*“... capable of integrating voice, data, and low-grade video (64 Kb/s to 2 MB/s) on ...”* | **number, ...** |
| **mathematical** | Mathematical entity.  
*“Polynomial equations are used for representing semialgebraic sets.”* | **formula, equation, ...** |
| **paper component** | A component of the research paper.  
*“...some successful applications (Section 3)....”* | **section, subsection, ...** |
| **date** | A date.  
*“In 1938 Albert Einstein and Leopold Infeld wrote...”* | **sequence of digits, ...** |
<table>
<thead>
<tr>
<th>Concepts</th>
<th>Explanation &amp; Example</th>
<th>Lexical Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>objective</td>
<td>The general objectives, “...the natural focus is on the third of the above goals...”</td>
<td>goal, objective, ...</td>
</tr>
<tr>
<td>conceptual objective</td>
<td>This includes the objective of a domain concept. “The aim of the DECIMAL (DECision-Making in Libraries) Project is to produce...”</td>
<td>goal of conceptual entity, ...</td>
</tr>
<tr>
<td>focus</td>
<td>The general focus. “A key focus of the technical specification was ...”</td>
<td>focus, ...</td>
</tr>
<tr>
<td>conceptual focus</td>
<td>This includes de focus of domain concept. “The focus of the paper is ...”</td>
<td>focus of conceptual entity, ...</td>
</tr>
<tr>
<td>topic</td>
<td>The general topic. “These main topics are natural-language processing,...”</td>
<td>topic, theme, ...</td>
</tr>
<tr>
<td>conceptual topic</td>
<td>the topic of the paper, paper component, study. “...a particularly important topic of study, and an issue that...”</td>
<td>topic of article, ...</td>
</tr>
<tr>
<td>introduction</td>
<td>Introducing information. “a brief introduction to this problem-solving paradigm...”</td>
<td>introduction, ...</td>
</tr>
<tr>
<td>overview</td>
<td>An overview. “...give a brief overview of the AI and robotics research performance...”</td>
<td>overview, ...</td>
</tr>
<tr>
<td>survey</td>
<td>A survey. “...Ameritech Library Services collaborated on a survey of electronic services...”</td>
<td>survey, ...</td>
</tr>
<tr>
<td>development</td>
<td>A development. “...the AI approach has prevailed in the implementation of high-level planning...”</td>
<td>development, ...</td>
</tr>
<tr>
<td>comparison</td>
<td>A comparison. “Aamodt’s comparison of knowledge intensive CBR methods...”</td>
<td>comparison, ...</td>
</tr>
<tr>
<td>analysis</td>
<td>An analysis. “To enable such an analysis of problem solving and learning, ...”</td>
<td>analysis, ...</td>
</tr>
<tr>
<td>presentation</td>
<td>A presentation. “...is demonstrated through the presentation of two current CBR systems...”</td>
<td>presentation, ...</td>
</tr>
<tr>
<td>discussion</td>
<td>A discussion. “Our discussion focuses on the main ...”</td>
<td>discussion, ...</td>
</tr>
<tr>
<td>Concepts</td>
<td>Explanation &amp; Example</td>
<td>Lexical Items</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>definition</td>
<td>A definition.</td>
<td>definition, ...</td>
</tr>
<tr>
<td></td>
<td>&quot;...a logical definition of the abductive problem...&quot;</td>
<td></td>
</tr>
<tr>
<td>description</td>
<td>A description.</td>
<td>description, ...</td>
</tr>
<tr>
<td></td>
<td>&quot;...the description of a problem, that has been...&quot;</td>
<td></td>
</tr>
<tr>
<td>explanation</td>
<td>An explanation.</td>
<td>explanation, ...</td>
</tr>
<tr>
<td></td>
<td>&quot;...and an explanation of ground facts cannot lead to rules.&quot;</td>
<td></td>
</tr>
<tr>
<td>suggestion</td>
<td>A suggestion.</td>
<td>suggestion, ...</td>
</tr>
<tr>
<td></td>
<td>&quot;A number of suggestions are put in place for...&quot;</td>
<td></td>
</tr>
<tr>
<td>discovery</td>
<td>A discovery.</td>
<td>discovery, ...</td>
</tr>
<tr>
<td></td>
<td>&quot;...to guide the discovery of repetitive functional substructures in large structural databases.&quot;</td>
<td></td>
</tr>
<tr>
<td>situation</td>
<td>The situation.</td>
<td>situation, today, ...</td>
</tr>
<tr>
<td></td>
<td>&quot;the Austrian situation in the field of telecommunication infrastructure is far behind</td>
<td></td>
</tr>
<tr>
<td>example</td>
<td>An example</td>
<td>example, illustration, ...</td>
</tr>
<tr>
<td></td>
<td>&quot;One example is &quot;concept formation&quot; as a goal,...&quot;</td>
<td></td>
</tr>
<tr>
<td>advantage</td>
<td>Advantage</td>
<td>advantage, asset, ...</td>
</tr>
<tr>
<td></td>
<td>&quot;...the advantages of CBR...&quot;</td>
<td></td>
</tr>
<tr>
<td>conclusion</td>
<td>Conclusion of the paper.</td>
<td>conclusion, ...</td>
</tr>
<tr>
<td></td>
<td>&quot;Our conclusion was that simple and local transformations...&quot;</td>
<td></td>
</tr>
<tr>
<td>summary</td>
<td>Summary of information.</td>
<td>summary, ...</td>
</tr>
<tr>
<td></td>
<td>&quot;...a summary of the fundings of the research phase.&quot;</td>
<td></td>
</tr>
</tbody>
</table>
# Annexe D

## Relations

<table>
<thead>
<tr>
<th>Relations</th>
<th>Explanation &amp; Example</th>
<th>Lexical Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>make known</td>
<td>Introducing the topic of the paper. “In this paper we present...”</td>
<td>describe, expose, ...</td>
</tr>
<tr>
<td>open</td>
<td>Opening. “...starts from a point-based metric system and gives a construction of...”</td>
<td>begin, start, ...</td>
</tr>
<tr>
<td>close</td>
<td>Closing. “The paper concludes with observations on the potential...”</td>
<td>conclude, close, ...</td>
</tr>
<tr>
<td>show</td>
<td>Identifying graphical material. “Figure 1 illustrates the concept...”</td>
<td>see, show, ...</td>
</tr>
<tr>
<td>summarize</td>
<td>Summarizing. “This analysis summarizes some of the work...”</td>
<td>sum up, ...</td>
</tr>
<tr>
<td>study</td>
<td>Studying a topic. “The possible reasons for this are explored as well as...”</td>
<td>analyze, examine, ...</td>
</tr>
<tr>
<td>investigate</td>
<td>Investigating. The phase transition in binary constraint satisfaction problems, i.e., ..., is investigated.</td>
<td>investigate, ...</td>
</tr>
<tr>
<td>situation</td>
<td>Identifying the situation. Genetic algorithms, the best known of the variants in the US, model evolution at the level of gene propagation.</td>
<td>know, ...</td>
</tr>
<tr>
<td>need</td>
<td>Identifying need. “In order to understand the French situation it is necessary to describe...”</td>
<td>to be a necessity, ...</td>
</tr>
<tr>
<td>experiment</td>
<td>To do experiments. “Experiments were done in which...”</td>
<td>do experiment, ...</td>
</tr>
<tr>
<td>discover</td>
<td>Discovering. “Significant reduction were discovered...”</td>
<td>determine, discover, ...</td>
</tr>
<tr>
<td>Relations</td>
<td>Explanation &amp; Example</td>
<td>Lexical Items</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>infer</td>
<td>Inferring. “The combination of the two methods...has been proven...”</td>
<td>prove, infer, ...</td>
</tr>
<tr>
<td>problem</td>
<td>Identifying problem. “... the main control problems that arise when ...”</td>
<td>arise, complicate, ...</td>
</tr>
<tr>
<td>solution</td>
<td>Identifying solution. “... it overcomes many of the past barriers to ...”</td>
<td>overcome, solve, ...</td>
</tr>
<tr>
<td>create</td>
<td>Bringing into being. “A new generation of sensor-rich, massively distributed, autonomous systems are being developed ...”</td>
<td>build, complete, ...</td>
</tr>
<tr>
<td>interest</td>
<td>Express interest. “We are concerned with the production of...”</td>
<td>address, concern, ...</td>
</tr>
<tr>
<td>focus</td>
<td>Identifying the focus. “We have focussed our development in two areas...”</td>
<td>focus on, ...</td>
</tr>
<tr>
<td>objective</td>
<td>Identifying the objective. “Its charter is to to perform research and development in advanced information technology...”</td>
<td>aim, ...</td>
</tr>
<tr>
<td>explain</td>
<td>Explaining. “The accuracy of a prediction based on the expected number of solutions is discussed ...”</td>
<td>discuss, explain, ...</td>
</tr>
<tr>
<td>opinion</td>
<td>Making a judgement. “A class of sparse problems ... is considered ...”</td>
<td>consider, believe, ...</td>
</tr>
<tr>
<td>argue</td>
<td>Argumenting. “It is argued that for most uses which are made of...”</td>
<td>argue, give argument, ...</td>
</tr>
<tr>
<td>comment</td>
<td>Commenting. “Notes the continuing explosion of information...”</td>
<td>mention, note, ...</td>
</tr>
<tr>
<td>suggest</td>
<td>Suggesting. “Specifically it is recommended that...”</td>
<td>suggest, ...</td>
</tr>
<tr>
<td>evidence</td>
<td>Giving evidence. “Evidence for the applicability of a conversational approach to 'information retrieval interaction' is initially provided by...”</td>
<td>to be evident, ...</td>
</tr>
<tr>
<td>conclude</td>
<td>Concluding. “The second conclusion to be drawn is that...”</td>
<td>conclude, ...</td>
</tr>
<tr>
<td>relevance</td>
<td>Identifying relevance. “Combinatorial and geometric computing is a core area of computer science”</td>
<td>to be central, ...</td>
</tr>
<tr>
<td>Relations</td>
<td>Explanation &amp; Example</td>
<td>Lexical Items</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>define</td>
<td>Defining. “Azuma (1997) has defined augmented reality systems as...”</td>
<td>define, to be, ...</td>
</tr>
<tr>
<td>describe</td>
<td>Describing. “The classical generative planning process consists of a search...”</td>
<td>compose, form, ...</td>
</tr>
<tr>
<td>essential</td>
<td>Identifying essentiality. “It is essential that all information staff...”</td>
<td>to be essential, ...</td>
</tr>
<tr>
<td>advantage</td>
<td>Identifying advantage. “...simulated annealing and evolutionary programming outperform back propagation.”</td>
<td>to have advantage, ...</td>
</tr>
<tr>
<td>use</td>
<td>Identifying usefulness. “The analytical tool which is used for this purpose is...”</td>
<td>apply, employ, ...</td>
</tr>
<tr>
<td>identify</td>
<td>Characterizing entity. “...a new algorithm called OPT-2 for optimal pruning...”</td>
<td>contain, classify, call, ...</td>
</tr>
<tr>
<td>exemplify</td>
<td>Exemplifying. “Other examples of robotic systems using fast vision tracking are also presented...”</td>
<td>to be example, ...</td>
</tr>
<tr>
<td>elaborate</td>
<td>Elaborating. “This properties allow us to use MT to express and prove tactics...”</td>
<td>allow, contribute, ...</td>
</tr>
<tr>
<td>effective</td>
<td>Identifying effectiveness. “Our algorithm is effective for...”</td>
<td>to be effective, ...</td>
</tr>
<tr>
<td>positive</td>
<td>Identifying positiveness. “...the workstations are promising”</td>
<td>to be positive, ...</td>
</tr>
<tr>
<td>novel</td>
<td>Identifying novelty. “The possibility of this new computing paradigm seems to rest on...”</td>
<td>to be new, ...</td>
</tr>
<tr>
<td>practical</td>
<td>Identifying practicality. “V&amp;V methodologies is a practical...”</td>
<td>to be practical, ...</td>
</tr>
<tr>
<td>perform</td>
<td>Doing. “...genetic programming techniques optimized for performing symbolic regressions.”</td>
<td>perform, ...</td>
</tr>
</tbody>
</table>
Annexe E

Types of Information in Selective Analysis

E.1 Indicative Types

**Topic of Document:** the author explicitly marks the topic of the document. This is identified by the presence of verbs of the make know relation, and concepts like the author or the research paper in first or last sections of the document.
**Ex.:** In *this paper we have presented* a more efficient distributed algorithm which construct a breadth-first search tree in an asynchronous communication network.

**Possible Topic:** some information is or will be presented. This is identified by the presence of verbs of the make know relation, in the passive form in first or last sections of the document.
**Ex.:** The sampling procedure *is described*, in which queries obtained from each library were broadly categorized by image content, identification and accessibility.

**Topic of Section:** the author explicitly marks what will be presented or discussed in a section of the document. This type of information is identified by the presence of the make know relation and the concept paper component.
**Ex.:** *Section 2 gives* the basic definitions for relation algebras and their representations together with some properties of representations.

**Closing:** the author closes the paper. This is marked by the close relation.
**Ex.:** *We conclude with* a summary of our contributions and a presentation of the implications of our results.

**Opening:** The author opens the paper. This is marked by the open relation.
**Ex.:** *We start with* economic theories of organization as the foundation of our analysis.
Goal of Conceptual Entity: the explicit mention of the objective of a domain concept (conceptual objective) or the domain concept (i.e. research paper, etc.) and the objective relation.
Ex.: The purpose of the experiments described here was to provide empirical based information regarding the imposition of a syntax and regarding the provision of visual prompts and feedback which could be used to guide decisions on design options for ASR interfaces.

Focus of Conceptual Entity: the center of attention of a domain concept identified by the presence of the concept conceptual focus or by the focus relation and a domain concept (i.e. author, etc.).
Ex.: Currently work is focused on the temporal aspects of the spatio-temporal reasoning techniques to be applied to GIS.

Author Development: the explicit mention of a development of the author. We identify this information by the co-occurrence of the author concept and create relation.
Ex.: As part of the UK Electronic Libraries programme, the authors have developed a simple decision support tool which allows a library manager to compare the total cost of acquiring a given item of information from each of a number of different sources.

Development: something was developed that could eventually be relevant. This is identified by the create relation.
Ex.: The syllabus has been designed to offer the Information Engineering Professional (IEP) opportunities to extend and develop their knowledge and skills.

Inference: some “conclusion” has been obtained from the research work. This is identified by the occurrence of research concepts such as work, result, experiment and the infer relation.
Ex.: Results shown that domain specific knowledge improves the search for substructures and enable greater data compression.

Author Interest: the authors refer to their interests. We identify that information by co-occurrence of the author concept and the interest relation.
Ex.: Finally we address the issue of scalability of structure discovery using Subdue.

Other’s Interest: some one (or something) shows interest in something. This is identified by the interest relation.
Ex.: The CBR paradigm addresses issues of experience-based design where experience is strong but the domain model weak, or poor formalized.
Problem: an explicit reference to a problem by the presence of the problem concept or relation.
Ex.: Lack of information skills in the majority of organizations

Solution: an explicit reference to a solution by the concept or relation solution.
Ex.: In this paper we propose a metatheory, MT, which represents the computation which implements its object theory, OT, and the computation which implements deduction in OT.

Topic: a topic is explicitly stated with the concept topic or conceptual topic and the define relation.
Ex.: The subject of this paper is the concept of descriptor equivalence and its two sub-concepts, dictionary equivalence and indexing equivalence.

Introduction of Entities: some entities introduced in the first section of the document, identified by the define or identify relations.
Ex.: The Danish Library Centre (DBC) is a central organization with Danish library and information sector.

Identification of Acronyms: an acronym and its expansion appear together in the same sentence.
Ex.: ...Distributed queue dual bus (DQDB) networks...

Signaling of Information: the author points to some important information appearing in some table or figure. This is identified by the relation show and the structural concept.
Ex.: Figure 1 illustrates the concept behind the proposed ATM network.

Signaling Concept: the author points to specific concepts such as introduction, presentation, overview, analysis, development, summary, etc.
Ex.: After a brief introduction to anytime computation, I outline a wide range of existing solutions to the meta level control problem and describe that is aimed at increasing the applicability of anytime computation.

Experiments: the mention of the experiments identified by the experiment concept.
Ex.: Experiment 1 studies the use of a syntax in a purely auditory interface.

Experimentation: experiments have been done. This is marked by the experiment concept and the experiment relation.
Ex.: Quality tests were completed with tests aimed to evaluate COES ability to detect errors and its performance.
Methodology: the mention of the methodology identified by the method concept.
Ex.: The combination of the two methods, hierarchical classification and Case-based reasoning, has been proven to be very successful in this specific textile application.

Author Study: the author refers to the study of some concept. This is identified by the author concept and the study relation.
Ex.: We analyse the complexity of our algorithm.

Explaining: Explanations are given. This is identified by the explain relation.
Ex.: I will discuss why they should consider these issues: explicit transfer of control, conflict analysis and reports, and autonomy hierarchy design.

Commenting: Comments are made. This is marked by the comment relation.
Ex.: I refer, of course, to the notion of Web-based computing with Java applets.

Giving evidence: Evidence is given. This is marked by the evidence relation.
Ex.: We first note that, when followed, the process leads to usable, useful, likeable computer systems and applications.

Object of Study: the mention of something that is (or was) studied identified by the study relation.
Ex.: In their current form, evolutionary computations are studied and applied in three standard formats, each distinguished by what level of granularity the algorithm models evolution.

Need for Research: an explicit mention of some need marked by the need relation or the statement of a research question. The information usually appeared in the first section.
Ex.: In order to understand the French situation it is necessary to describe the strategy presently followed by France Telecom, which remains the inheritor of the policy of the 1980s.

Actual Situation: this information is usually stated in the introduction of the document using the situation relation or concept.
Ex.: Currently, most geometric objects (curves and surfaces) are formulated in terms of polynomial equations, thereby reducing many application problems - such as boundary computations - to manipulating polynomial systems.

Hypothesis: the statement of the hypothesis. This is identified by the hypothesis concept.
Ex.: As in Damper and Wood's work, the hypothesis under test is that the observed superiority of speech over keying in Pook's experiments arises from a methodological flow.
** Authors' Opinion:** the authors express their opinion. We identify this information by the co-occurrence of the concept *author* and the relation *opinion*.

**Ex.:** The scheme is based on where the sort from an object coordinates to screen coordinates occurs, which *we believe* is fundamental whenever both geometry processing and rasterization are performed in parallel.

**Authors' Discovery:** the authors have made a discovery. This is marked by the concept *author* and the *discover* relation.

**Ex.:** *I determined* the communication costs for classes of parallel algorithms by considering their inherent communication requirements.

**Authors' Argument:** the authors make an argument. We identify this using the concept *author* and the *argue* relation.

**Ex.:** *We have argued* that the problem of invisibility and the related problem of complexity are caused by information barriers.

**The Authors Demonstration:** the authors deduce something. We identify this by the presence of the *author* concept and the *infer* relation.

**Ex.:** *We have demonstrated* that the service offered by ATM layer in the B-ISDM Practical reference model is equivalent to the service offered by the OSI physical layer.

**Investigation:** something is being investigated and this is marked by the *investigate* relation.

**Ex.:** Next, the network reliability is *investigated*, both for degradation in throughput and loss in connectivity.

**Suggestions:** something is suggested. This is identified by the relation *suggest* or by the concept *suggestion*.

**Ex.:** *Suggests* that perhaps online searching should be learned by watching, assisting, and being corrected by a master searcher.

**Conclusions:** something is concluded marked by the concept *conclusion* or by the relation *conclude*.

**Ex.:** *Our conclusion* was that simple and local transformations can be automatized or semi-automatized, depending whether additional information is not needed, while global transformations are difficult to automatize.

**Summarization of Information:** information is being summarized. This is identified by the concept *summary* or the relation *summarize*.

**Ex.:** It closes with an outline of future directions for SIP development, a report on current implementation status and a *summary* of the specific improvements offered by SIP over IP.
E.2 Informative Types

Relevance of Entity: the mention of a relevant entity by referring to the relevance relation.
Ex.: Indexing a content rich in semantic is a key to interpreting a design problem and the selection of an adaptation strategy.

Goal of Entity: the explicit mention of the objective of a non conceptual entity. This is marked by the objective concept or relation.
Ex.: The goal of CCAD is to support exploratory design, while keeping the user central to the design activity.

Essential of Entity: the mention of the essential of an entity, using the essential relation.
Ex.: The essential element of success is the integration of literacy efforts into the overall library operations.

Positiveness of Entity: the mention of a positive aspect of an entity. This is identified by the positive relation.
Ex.: It frees the programmer to concentrate on parallel algorithms instead of low-level implementation details, and it yields good performance.

Usefulness of Entity: a reference to the applicability of an entity. This is marked by the use or perform relation or by the application concept.
Ex.: Our current model is easy to use, and it supports several higher level programming constructs in several languages, including micro tasking in C and Fortran and multitasking in Ada and C++.

Effectiveness of Entity: a reference to the goodness of an entity. We identify this information using the effective relation.
Ex.: Our compositing algorithm is particularly effective for massively parallel processing.

Description of Entity: an entity is being described. This is identified by the describe relation.
Ex.: The algorithm is based on dynamic programming.

Definition of Entity: an entity is being defined. This is marked by the define relation.
Ex.: To this end an effect language is defined which connects the designer's description with the implemented functionality of the application.
E.2. INFORMATIVE TYPES

**Advantage of Entity:** an explicit mention to the advantage of an entity by using a **advantage** concept or relation.  
**Ex.:** *The advantage* of our method is obvious.

**Practicality of Entity:** the explicit mention of the actual application of an entity. This is identified by the **practical** relation.  
**Ex.:** This article details simple mechanisms for introducing hierarchy into the inter-domain in routing systems, making it **practical** to route a truly large Internet.

**Novelty of Entity:** some new entity is introduced marked by the **novel** relation.  
**Ex.:** In this paper a **new** algorithm called OPT-2 for optimal pruning of decision trees is introduced.

**Elaboration of Entity:** some very specific elaborations of entities using the **elaborate** relation.  
**Ex.:** PiP **provides** new features, such as provider selection.

**Exemplification:** something is being exemplified. This is marked by the relation **illustrate** or the **example** concept.  
**Ex.:** *We illustrate* TPQN’s capabilities with a performance analysis of a real-time, multitasking scheduler that had been previously implemented on top of SunOS on Sun-3 workstation.
Annexe F

Parsed Segment Fragments

This activity (Dete A+ N+)
(sem, activity), (conceptual, dc), (concept, activity),
(anaphoric, definite), (syncat, gn), (gntype, GN2),
(string, [This, activity]), (canon, [activity]), (DeteType, dtem),
(Typ, def), (Nbr, plur)

The Automatic Control Department (Dete A+ N+)
(sem, department), (conceptual, dc), (concept, institution), (syncat, gn),
(gntype, GN4), (string, [The, Automatic, Control, Department]),
(canon, [automatic, control, department]), (DeteType, dart), (Typ, def),
(Nbr, sing)

two co-operative robots (Quan A+ N+)
(sem, robot), (syncat, gn), (gntype, GN10), (string, [two, co-operative, robots]),
(canon, [co-operative, robot]), (QuanCla, num), (Nbr, plur)

a tele-operation station (Dete A+ N+)
(sem, station), (syncat, gn), (gntype, GN4), (string, [a, tele-operation, station]),
(canon, [tele-operation, station]), (DeteType, dart), (Typ, ind),
(Nbr, plur)

urban infrastructures (A+ N+)
(sem, infrastructure), (syncat, gn), (gntype, GN3),
(string, [urban, infrastructures]), (canon, [urban, infrastructure]),
(Nbr, plur)
the long term goal (Dete A+ N+)
(sem, goal), (conceptual, dc), (concept, goal), (syncat, gn),
gntype, GN4), (string, [the, long-term, goal]), (canon, [long-term, goal]), (DeteType, dart), (Typ, def), (Nbr, sing)

This paper (Dete paper)
(syncat, gn), (conceptual, dc), (concept, research, paper), (type, author),
(id, paper), (string, [This, paper]), (Nbr, sing)

Nuclear power plant steam generator inspection (A+ N+)
(sem, inspection), (syncat, gn), (gntype, GN3), (string, [Nuclear, power, plant, steam, generator, inspection]), (canon, [nuclear, power, plant, steam, generator, inspection]), (Nbr, sing)

SIROIN (NomP+)
(sem, SIROIN), (syncat, gn), (gntype, GN5), (string, [SIROIN]),
(canon, [SIROIN]), (Nbr, sing)

an innovative solution (Dete A+ N+)
(sem, solution), (conceptual, dc), (concept, solution),
(quality, novelty+), (syncat, gn), (gntype, GN4), (string, [an, innovative, solution]), (canon, [innovative, solution]),
(DeteType, dart), (Typ, ind), (Nbr, sing)

the climbing and walking robots (Dete Sequence, gn)
(sem, robot), (DeteType, dart), (Typ, def), (syncat, gn), (gntype, GN*),
(string, [the, climbing, and, walking, robots]), (canon, [climb, and, walk, robot]), (Nbr, plur)

Inove (1994) (NomP (Year))
(syncat, gn), (gntype, ref), (reftype, 'Ref1'), (conceptual, dc),
(concept, reference), (researcher, 'Inove'), (year, 1994),
(string, ['Inove', '(', '1994', ']'), (Nbr, nil)

briefly outlines (Adv Verb)
(conceptual, dr), (pred, outline), (type, none), (relation, 'to make know'),
(syncat, gv), (Nbr, plur), (tense, sim-pre), (voice, active),
(time, pres), (string, [briefly, outlines]), (canon, outline),
(adv, briefly)
was approached ( was Verb )
(conceptual,dr), (pred,approach), (type,none), (relation,approach problem), (syncat,gv), (Nbr,sing), (tense,sim_pas), (voice,pasive),
(time,pas), (string,[was,approached]), (canon,approach)

is implemented ( is Verb )
(conceptual,dr), (pred,implement), (type,none), (relation,to bring into being), (syncat,gv), (Nbr,sing), (voice,pasive), (tense,sim_pre),
(time,pre), (string,[is,implemented]), (canon,implement)

demonstrated ( Verb )
(conceptual,dr), (pred,demonstrate), (type,none), (relation,to describe), (relation,to elaborate entity), (relation,to identify result), (relation,to infer), (syncat,gv), (tense,sim_pas),
(voice,active), (time,pas), (string,[demonstrated]),
(canon,demonstrate)

is shown ( is Verb )
(conceptual,dr), (pred,show), (type,lex), (relation,to describe),
(relation,to identify result), (relation,to infer), (relation,to make know), (relation,to show graphical material), (syncat,gv), (Nbr,sing),
(voice,pasive), (tense,sim_pre), (time,pre), (string,[is,shown]),
(canon,show)

mandatory ( A+ )
(conceptual,da), (quality,necessary), (syncat,A+),
(string,[mandatory]), (canon,[mandatory])

Figure 3 ( Figure Quan )
(syncat,gn), (conceptual,dc), (concept,structural), (type,figure),
(id,3), (string,[Figure,3])
Annexe G

Instruction Booklet for the Evaluators

G.1 Overview

My research is concerned with the generation of abstracts of technical articles by automatic means. I am implementing a program which produces indicative abstracts and I am interested in evaluating how this first implementation works. In order to produce an abstract, the automatic system extracts information from the source document (technical article) and presents it to the reader.

G.2 Evaluation

The objective of this experiment is to evaluate the abstract produced by the system. Two aspects are addressed: first, if the abstracts the system produces are good in indicating the content of the source document, and second, how acceptable they are.

The source documents used in this evaluation were obtained from the journal “Industrial Robot.” The abstracts to evaluate were produced by two different summarization systems or were originally published in the journal.

You will receive a form with a number of abstracts to evaluate in separate pages. You will not be informed about the method used to produce the abstract. Each abstract is to be evaluated in two dimensions: content and quality.

1. Content
   Along with the abstract to evaluate, you will receive 5 lists of keywords which are meant as content indicators. Those lists were obtained from the journals where the source documents were published. One of the lists corresponds to the abstract to evaluate. I would like you to read carefully the abstract and check the list of keywords that best matches the content of the abstract (you can check more than one or none at all).

2. Text quality
   In order to evaluate the quality of the text I would like you to provide an acceptability
score between 0-5 for the abstract (0 for unacceptable and 5 for acceptable). Abstracts with scores ranging between 2.5 and 5 will be considered acceptable.

Some of the criteria you could use in order to evaluate the texts as acceptable are Rowley (1982):

- good spelling and grammar;
- clear indication of the topic of the source document;
- impersonal style;
- one paragraph;
- conciseness (no unnecessary references to the source document);
- readable and understandable;
- acronyms are presented along with their expansions; and
- other criteria that you might consider important as an experienced reader of abstracts of technical documents.

Please note that you might consider an abstract acceptable even if it contains minor errors. You can provide comments at will to help me understand the score you gave the text indicating which criteria the abstract fulfills or fails to fulfill.

Note that in these experiments I am by no means evaluating you as a reader but the collected data will be used to evaluate my system. The results of this experiments will be used for research purposes and all the data about the subjects will be kept confidential.

Please, take your time to accomplish this task and do not forget to complete the form with the information about your skills in English. In order to know how hard it was for you to accomplish this task, I would appreciate if you provide the time it took you to complete the whole form. I would appreciate any comments about this experiment. Please use the last page in the form for that purpose.

My eternal gratitude for your participation.

Horacio Saggion

e-mail: saggion@iro.umontreal.ca
Annexe H

Informed Consent Form to Participate in Research

**Purpose**: This research is concerned with the generation of abstracts of technical articles by automatic means. This investigation is expected to contribute to the field of automatic abstracting by proposing a new method of text summarization and by demonstrating the viability of such a method.

**Procedures**: In this experiment you will be asked to read and evaluate abstracts produced by humans and by two different automatic abstracting systems. The source documents used to produce the abstracts are technical documents from the journal “Industrial Robots.” You will not be informed about the method that was used to produce each abstract (either automatic or manual). The abstracts are to be evaluated in content and text quality. It is expected that the overall evaluation will take a maximum of 60 minutes including explanations by the researcher.

For content evaluation you will be asked to chose a content indicator for the abstract out of a list of 5 content indicators given along with each abstract. One of the content indicators corresponds to the abstract. The other four were randomly chosen from the journal where the source documents were published. You can chose more than one or none at all.

For text quality, you will be asked to provide an acceptability score between 0 and 5 (0 means “unacceptable” and 5 means “acceptable”) based on the following criteria:

- good spelling and grammar;
- clear indication of the topic of the source document (topical sentence);
- impersonal style;
- one paragraph;
- conciseness (no unnecessary references to the source document);
- readable and understandable;
- acronyms are presented along with their expansions; and
other criteria that you might consider important as an experienced reader of abstracts of technical documents.

Conditions of Participation: It is very important to note that the researcher wants to evaluate and compare the results of an automatic system and that your abilities or linguistic skills are not evaluated. Your responsibility in this experiment is to complete the evaluation form and give it to the researcher. This experiment does not entail any risk for the participants.

The present consent form will be stored in a file separate from the evaluation data. Only information about your reading and comprehension skills in English is required. The data collected in the forms will be used to calculate averages of text quality and indicativeness for the three methods.

The results of this experiment will be used for research purposes and all the nominal information about the subjects will be kept confidential. The results will be published as part of a Ph.D. dissertation and eventually disseminated in research papers.

This is to state that I agree to participate in the research project entitled Evaluation of Automatic Abstracting Systems conducted by Horacio Saggion at Département d’informatique et de recherche opérationnelle (DIRO), University of Montreal, and supervised by Professor Guy Lapalme.

- I have been informed that the purpose of the research is to investigate the viability of a new method of automatic abstracting of texts.
- I understand the purpose of this study and know about the risks, benefits and inconveniences that this research project entails.
- I understand that I am free to withdraw at anytime from the study without any penalty or prejudice.
- I understand how confidentiality will be maintained during this research project.
- I understand the anticipated uses of data, specially with respect to publication, communication and dissemination of results.

I have carefully studied the above and understand my participation in this agreement. I freely consent and voluntarily agree to participate in this study.

- Name (please print):
- Signature:
- Date:
Annexe I

Questionnaire Sample for Evaluation of Indicativeness

Group VI

– Please, provide the following information about your skills in English:

<table>
<thead>
<tr>
<th>Reading and Comprehension</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
</tr>
<tr>
<td>Acceptable</td>
<td></td>
</tr>
<tr>
<td>with difficulty</td>
<td></td>
</tr>
<tr>
<td>not at all</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English is your...</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>First Language</td>
<td></td>
</tr>
<tr>
<td>Second Language</td>
<td></td>
</tr>
<tr>
<td>Other (specify)</td>
<td></td>
</tr>
</tbody>
</table>
ABSTRACT 21:

High speed arc welding
Welding current and spatter
Past methods for welding control
High speed systems
MOTOPAC-WH200 is an arc welding robot package specifically designed for high-speed welding.
Welding thin plates
Welding intermediate thickness plates
Welding thick plates
By achieving high-speed welding, the arc welding times are shortened by half.
Figure 2 High speed arc welding technology.
Figure 5 Servotorch.

CONTENT

Please check (X) the box next to the content indicator you think corresponds to the text above. You can check more than one or none at all.

<table>
<thead>
<tr>
<th>Content Indicators</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robots, Service, Telexistence, Teleoperation, VR, Virtual Reality</td>
<td></td>
</tr>
<tr>
<td>Asbestos, Insulation, Robotics</td>
<td></td>
</tr>
<tr>
<td>Control, Language, Robots</td>
<td></td>
</tr>
<tr>
<td>Robots, Welding, Automotive</td>
<td></td>
</tr>
<tr>
<td>Language, Neural networks, Robots</td>
<td></td>
</tr>
</tbody>
</table>

QUALITY

Please give your score to the above text according to the criteria specified in the instruction booklet.

| Your Score for the Text (0-5) : .......... |

Please include here your comments about the text :
ABSTRACT 23:

Many research efforts have turned to sensing, and in particular computer vision, to create more flexible robotic systems. Computer vision is often required to provide data for the grasping of a target. Using a vision system for grasping of static or moving objects presents several issues with respect to sensing, control, and system configuration. This paper presents some of these issues in concept with the options available to the researcher and the trade-offs to be expected when integrating a vision system with a robotic system for the purpose of grasping objects. The paper includes a description of our experimental system and contains experimental results from a particular configuration that characterize the type and frequency of errors encountered while performing various vision-guided grasping tasks. These error classes and their frequency of occurrence lend insight into the problems encountered during visual grasping and into the possible solution of these problems.

CONTENT

Please check (X) the box next to the content indicator you think corresponds to the text above. You can check more than one or none at all.

<table>
<thead>
<tr>
<th>Content Indicators</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robots, Service, Telexistence, Teleoperation, VR, Virtual Reality</td>
<td></td>
</tr>
<tr>
<td>Control, Language, Robots</td>
<td></td>
</tr>
<tr>
<td>Materials handling, Robotics, Sensors, Systems, Tracking</td>
<td></td>
</tr>
<tr>
<td>Robots, Health care</td>
<td></td>
</tr>
<tr>
<td>Metrology, Photogrammetry, Robots</td>
<td></td>
</tr>
</tbody>
</table>

QUALITY

Please give your score to the above text according to the criteria specified in the instruction booklet.

[Your Score for the Text (0-5) : ........]

Please include here your comments about the text :
ABSTRACT 26:

The big-on-asbestos (BOA) system is a mobile pipe-external robotic crawler used to remotely strip and bag asbestos-containing lagging and insulation materials (ACLM) from various diameter pipes in (primarily) industrial installations. Steam and process lines within the Department of Energy (DoE) weapons complex warrant the use of a remote device due to the high labor costs and high level of radioactive contamination, making manual removal extremely costly and highly inefficient. Currently targeted facilities for demonstration and remediation are Fernald in Ohio and Oak Ridge in Tennessee.

CONTENT

Please check (X) the box next to the content indicator you think corresponds to the text above. You can check more than one or none at all.

<table>
<thead>
<tr>
<th>Content Indicators</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language, Neural networks, Robots</td>
<td></td>
</tr>
<tr>
<td>Asbestos, Insulation, Robotics</td>
<td></td>
</tr>
<tr>
<td>Robots, Service, Telexistence, Teleoperation, VR, Virtual Reality</td>
<td></td>
</tr>
<tr>
<td>Robots, Health care</td>
<td></td>
</tr>
<tr>
<td>Control, Language, Robots</td>
<td></td>
</tr>
</tbody>
</table>

QUALITY

Please give your score to the above text according to the criteria specified in the instruction booklet.

Your Score for the Text (0-5): ..........
ABSTRACT 29:

Telexistence (tele-existence) is technology which enables a human being to have a real time sensation of being at a remote location, while giving the person the ability to interact with the remote and/or virtual environments. He or she can "telexist" (tele-exist) in a real environment where the robot exists or in a virtual environment that a computer has generated. It is also possible to telexist in a mixed environment of real and virtual, which is called augmented telexistence. The concept of telexistence, i.e. virtual existence in a remote or computer-generated environment, has developed into a national R&D scheme called R-Cubed (Real-time Remote Robotics). Based on the scheme the National R&D Project of "Humanoid and Human Friendly Robotics", Humanoid Robotics Project (HRP) in short, was launched in April 1998. This is an effort to integrate telerobotics, network technology and virtual reality into networked telexistence.

CONTENT

Please check (X) the box next to the content indicator you think corresponds to the text above. You can check more than one or none at all.

<table>
<thead>
<tr>
<th>Content Indicators</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robots, Service, Telexistence, Teleoperation, VR, Virtual Reality</td>
<td></td>
</tr>
<tr>
<td>Robots, Health care</td>
<td></td>
</tr>
<tr>
<td>Materials handling, Robotics, Sensors, Systems, Tracking</td>
<td></td>
</tr>
<tr>
<td>Asbestos, Insulation, Robotics</td>
<td></td>
</tr>
<tr>
<td>Robots, Welding, Automotive</td>
<td></td>
</tr>
</tbody>
</table>

QUALITY

Please give your score to the above text according to the criteria specified in the instruction booklet.

| Your Score for the Text (0-5): ........ |

Please include here your comments about the text:
ABSTRACT 31:

The McKibben muscle was invented in the 1950s by the American atomic physicist Joseph L. McKibben with the aim of motorizing an arm orthosis for the poliomyelitic daughter. The forgotten invention was rediscovered in the 1980s by the Japanese pneumatic tyre manufacturer, Bridgestone, who had remarked the similarity between the artificial muscle components and pneumatic tyre ones. Describes the McKibben muscle and analyses the performances of robot-arms driven by the artificial muscles. Shows the McKibben muscle, defined by the same parameters as the ones considered, performing an isotonic contraction. A basic controller was designed to test the robot including two levels: a point-to-point trapezoidal velocity profile trajectory generator and a joint closed-loop control. Shows original use of McKibben muscle as forearm and force change principle.

CONTENT

Please check (X) the box next to the content indicator you think corresponds to the text above. You can check more than one or none at all.

<table>
<thead>
<tr>
<th>Content Indicators</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robotics</td>
<td></td>
</tr>
<tr>
<td>Materials handling, Robotics, Sensors,</td>
<td></td>
</tr>
<tr>
<td>Systems, Tracking</td>
<td></td>
</tr>
<tr>
<td>Control, Language, Robots</td>
<td></td>
</tr>
<tr>
<td>Metrology, Photogrammetry, Robots</td>
<td></td>
</tr>
<tr>
<td>Calibration, Force sensing, Robots, Torque</td>
<td></td>
</tr>
<tr>
<td>sensing</td>
<td></td>
</tr>
</tbody>
</table>

QUALITY

Please give your score to the above text according to the criteria specified in the instruction booklet.

Your Score for the Text (0-5): ........

Please include here your comments about the text:
ABSTRACT 36:
Teleoperator slave - WAM design methodology between the master and slave. Independent roots of teleoperator masters and robot slaves. Now consider the design requirements of masters.
Master-slave symmetry
True to the notions of master and slave, many teleoperator designs enforce motion obedience on the slave without any channel for force-torque feedback to the master.
Figure 2 Single axis teleoperator with joint-torque-controllable slave illustrates this force-feedback symmetry for one axis.
Improved methods for designing slaves
Conventional robots retrofitted with force-torque sensors
Bandwidth
Force fidelity joints.
Borrowing elements from earlier master designs
WAM innovations remove limitations of cable drives
WAM design
Figure 1 Telerobotic system.
Figure 2 Single axis teleoperator with joint-torque-controllable slave.
Figure 3 Simple mechanical 50 :1 power transmission.

CONTENT
Please check (X) the box next to the content indicator you think corresponds to the text above. You can check more than one or none at all.

<table>
<thead>
<tr>
<th>Content Indicators</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robots, Welding, Automotive</td>
<td></td>
</tr>
<tr>
<td>Calibration, Force sensing, Robots, Torque sensing</td>
<td></td>
</tr>
<tr>
<td>Robots, Teleoperator</td>
<td></td>
</tr>
<tr>
<td>Control, Language, Robots</td>
<td></td>
</tr>
<tr>
<td>Materials handling, Robotics, Sensors, Systems, Tracking</td>
<td></td>
</tr>
</tbody>
</table>

QUALITY
Please give your score to the above text according to the criteria specified in the instruction booklet.

**Your Score for the Text (0-5): ........**

Please include here your comments about the text:
PLEASE INCLUDE HERE YOUR COMMENTS ABOUT THIS EVALUATION (YOU CAN PROVIDE THE TIME IT TOOK YOU TO ACCOMPLISH THIS TASK) :
Annexe J

Abstracts Produced by SumUM

We include for each document the abstract produced by SumUM and the abstract published by the journal in that order. We also provide word counts for the document, author abstract and automatic abstract.


Source Document: 5227 words
Author Abstract: 165 words (3.15%)
SumUM: 119 Words (2.30%)

Offline robot programming systems therefore require an accurate description of the workpieces and layout of the environment. Describes the walk-through programming approach and the algorithms to enable walk-through motion capabilities in industrial robots. A new man-machine interface that includes custom designed teach pendants and Graphical User Interface was developed. The gantry was designed to accommodate the 12m x 12m panel size requirements in Keppel FELS, a shipyard in Singapore. The handle was designed for ease in teaching and no interference during welding. A Graphical User Interface was developed to facilitate the selection of the welding parameters. The software module of each subsystem includes exception handling routines which ensure that the software will not be aborted abnormally leaving the system in an unknown state.

alternative approach, custom, dynamic parameters, ease, industrial robots, main pendant, massive gantry, motions, non-model-based approach, overall system, physical parameters, process parameters, robot, Robot teaching, robotic system, shipyards, dof robot, systems, typical panel, user, walk-through, welding parameters, welding systems

Automating the welding process for the shipbuilding industry is very challenging and important, as this industry relies heavily on quality welds. Conventional robotic welding systems are seldom used because the welding tasks in shipyards are characterised by non-standardised
workpieces which are large but small in batch sizes. Furthermore, geometries and locations of the workpieces are uncertain. To tackle the problem, a Ship Welding Robot System (SWERS) has been developed for the welding process. The main features of the SWERS include a special teaching procedure that allows the human user to teach the robot welding paths at a much easier and faster pace. In addition, operation of the system is made easier through a custom designed man-machine interface. Through this interface, only a few buttons need to be pressed to command the robot into different modes. Optimised welding parameters can be selected from a large database through a Graphical User Interface system.


Source Document: 2691 words  
Author Abstract: 108 words (4.00%)  
**SumUM:** 95 Words (3.50%)

The complex problem of inspection and maintenance of the steam generator in nuclear power plants was approached using previously gained expertise and, as a result, an innovative solution was achieved with the development of two co-operative robots, remotely controlled from a tele-operation station incorporating tele-presence. Outlines the developments carried out in robotic systems for hazardous environments in the department. Another interesting activity was the realization, within the framework of a EUREKA project, of a tele-manipulator for servicing a new concept of urban infrastructures. Shows the RIMHO walking robot and concept of IUI (Industrializable urban infrastructures).

**IUI, RIMHO, climbing robot, control station, control systems, department, different robotic systems, industrial robots, infrastructures, inspection, known and best referenced hazardous environment, robots, robotics system, subsequent maintenance, systems, tele-presence, wheeled mobile robot, whole control system**

Explains how the Automatic Control Department of the Instituto de Automatica Industrial (CSIC) in Madrid, Spain has been developing robots for over 15 years. This activity began in the 1980s with the realization of industrial robots and then the department focused its attention on the area of robots for hostile/hazardous environments. Describes several achievements in this field including a complex tele-operated system for steam generator inspection and maintenance in nuclear power plants; a tele-manipulator for servicing a new concept of urban infrastructures; a self-propelling climbing robot with magnetic feet; and a four-legged walking robot for hazardous environments.


Source Document: 3969 words  
Author Abstract: 131 words (3.30%)
Surgery traditionally involves making large incisions to get to the part of the patient that requires attention. It is therefore desirable to have a robotic system that can collaboratively perform endoscopic procedures with the surgeon, specifically performing certain tasks autonomously to reduce the strain on the surgeon, to remove the variability of surgeon’s training level, and to enhance the system efficiency by decreasing the operation time. Presents the design of a general-purpose computer controlled platform for minimally invasive surgery. The robotic tool contains two detachable portions. Shows block diagram for autonomous control.

autonomous suture operation, control, delicate operations, gripping tool, invasive surgery, MIS operations, platform, surgeon, suture tool, suturing tool, system, tool, usual PID control

Minimally invasive surgery (MIS) is a cost-effective alternative to the open surgery whereby essentially the same operations are performed using specialized instruments designed to fit into the body through several tiny punctures instead of one large incision. The EndoBots (Endoscopic Robots) described here are designed for collaborative operation between the surgeon and the robotic device. The surgeon can program the device to be operated completely manually, collaboratively where motion of the robotic device in certain directions is under computer control and in others under manual surgeon control, or autonomously where the complete device is under computer control. Furthermore, the robotic tools can be quickly changed from a robotic docking station, allowing different robotic tools to be used in an operation.


Source Document: 1609 words
Author Abstract: 20 words (1.25%)
SumUM: 18 Words (1.25%)

Explains the internal design of the Preci-Check measuring system; and also discusses the collaboration of hardware and software.

Preci-Check, system, temperature compensation systems

Describes the use of a robot as a measuring system for checking the manufacturing tolerance of complex 3-D assemblies.


Source Document: 3009 words
Author Abstract: 140 words (4.65%)  
**SumUM:** 17 Words (1.00%)  

Shows the RobotScript program; presents comparison of Robot Language Syntax; and also shows the universal robot controller.

**KAREL language, RobotScript, RobotScript program, computer language, computer programs, language, program, programming language, robot, robot language, robot programs, software programs**

The flexibility of a robot system comes from its ability to be programmed. How the robot is programmed is a main concern of all robot users. A good mechanical arm can be underutilized if it is too difficult to program. The introduction of the Universal Robot Controller (URC) has made the possibility of a standard, easy to use, robot programming language a reality. The URC is an open-architecture, PC-based robot controller. It will work with virtually any robot and gives the user increased flexibility and capabilities over the standard OEM controllers. The URC uses Windows NT as its operating system. The URC is the ideal platform for a universal robot programming language, RobotScript. It allows one robot language to run all robots in a factory.


Source Document: 2650 words  
Author Abstract: 60 words (2.26%)  
**SumUM:** 77 Words (2.90%)  

Presents a situation which needs calibration. The research aimed to develop a methodology which can calibrate the relative positioning inaccuracy on-line while the robot is carrying out the routine tasks without the need of mathematical models and complex off-line calibration procedures. Shows system hardware configuration for deriving major orientation errors with respect to the global co-ordinate frame and fine motion training diagram; presents neural network for on-line error compensation; and also illustrates the investigated example.

**example, methodology, models, orientation calibrations, proposed on-line calibration methodology**

Presents an on-line calibration methodology for robot relative positioning inaccuracy. This methodology eliminates the need for time-consuming off-line calibrations relying on accurate models and complicated procedures. To realize this methodology, a vision system, a 3D force/torque sensor, and control strategies involving Neural Networks (NNs) were incorporated with an industrial robot.

*Approaches for resolving dynamic IP addressing.* Schubert Foo. Internet Research : Electro-

Source Document: 4161 words
Author Abstract: 185 words (4.44%)
**SumUM:** 58 Words (1.40%)

Proposes a number of methods to overcome the dynamic IP addressing problem. It is concluded that the dynamic Domain Name System and the directory server look-up are the two best approaches for resolving dynamic IP addressing. Shows IP address resolution using exchange server and IP address resolution using directory service look-up; and also presents comparison of online methods.

E-mail address, E-mail approach, IP, IP address, WWW approach, Approaches, awkward IP, corresponding IP, current IP, directory servers, directory server look-up, directory service, directory service look-up approach, Domain Name System, dynamic IP, dynamic Domain Name System, exchange, exchange server, fixed IP, general servers, mail server, methods, name servers, online method, POP server, problem, servers, services, telephone look-up server, unique address, various default mail systems

Knowledge of the Internet Protocol (IP) address is essential for connection establishment in certain classes of synchronous distributed applications, such as Internet telephony and video-conferencing systems. A problem of dynamic IP addressing arises when the connection to the Internet is through an Internet service provider, since the IP address is dynamically allocated only at connection time. Proposes and draws a contrast between a number of generic methods that can be classified as online and offline methods for the resolution of dynamic IP addressing. Online methods, which include the World Wide Web, exchange server and the dynamic Domain Name System, are only effective when both the caller and recipient are logged on to the Internet. On the other hand, offline methods, which include electronic mailing and directory service look-up, provide an additional means to allow the caller to leave messages when the recipient is not logged on to the Internet. Of these methods, the dynamic Domain Name System and directory service look-up appear to be the best for resolving dynamic IP addressing.


Source Document: 4725 words
Author Abstract: 100 words (2.11%)
**SumUM:** 80 Words (1.70%)

Identifies in the study the key parameters of the LOM (The laminated object manufacturing) process, in terms of adequate bonding and cutting accuracy. The paper investigates the characteristics of the LOM process including an optimization of key system parameters,
the precision and accuracy of the LOM process, the dimensional stability of LOM parts, and the properties of LOM paper. The dimensional stability of LOM parts as a function of time was measured. Each part to be fabricated on LOM was designed using I-DEAS software.

LOM, LOM process, accuracy, compensation function, key process parameters, largest since dimensional stability, part, post treatment processes, precision, test parts, three-dimensional part

Laminated object manufacturing (LOM) is a rapid prototyping process where a part is built sequentially from layers of paper. Studied in the present paper are the precision and accuracy of the LOM process and the dimensional stability of LOM parts. The process was found to exhibit both constant and random sources of error in the part dimensions. The dimensional error was the largest normal to the plane of the paper, exacerbated by the moisture absorption and subsequent swelling. The key process parameters were identified and optimized for sufficient bonding and cutting accuracy.


Source Document: 4622 words
Author Abstract: 80 words (1.73%)
SumUM: 118 Words (2.55%)

While the use of stereolithography for producing test and small run molds has been well established, the life of the molds is limited. The parts were built, cleaned, and UV post-cured, the hardness was measured. Presents then the results. The DSC (Calorimeter) tests indicate a difference in the degree of cure between the inside material and the outside material of stereolithography samples. The results indicate that the surface hardness increases with time, and that after two days, all of the parts reach the same hardness. The test also showed that for the resins tested, there were no significant differences in the degree of cure. The DSC scan of the sample is shown. Presents cure behavior of stereolithography polymers.

DSC tests, SLA material, UV post-cured, cure, cured parts hardness, inside and the outside material, inside material isothermal test, liquid resin, material, material degradation tests, molds, outside, outside material, parts, results, Samples, spectroscopy results, stereolithography, stereolithography material, test

A common procedure for processing stereolithography epoxy injection molds includes a one hour post-cure in a UV chamber. This research investigates the degree of cure achieved in the UV chamber and the degree of cure achieved by heating in a thermal oven. It is hypothesized that a more fully cured mold is harder and hence will produce more parts before failure. This research investigates various post-cure processes and suggests a post-cure strategy to achieve this end.

Source Document: 2674 words  
Author Abstract: 96 words (3.60%)  
**SumUM**: 28 Words (1.04%)  

GENERIS has a multi-machine architecture that facilitates the manufacturing process control of a real workshop containing several robots, auxiliary devices and process sensor systems. Shows GENERIS functional model.

**ABB-SPEEDY robot, CREA-AMADA EUROSCARA robot, GENERIS, Vx-Works operating system, multi-machine, real-time operating system, robots, SMART S2-COMAU robot, software systems, system**

This paper presents an overview of the main features of GENERIS, a generalised software control system for industrial robots developed by the EC-JRC and financed by the EC-DGXIII-D1. After an introduction reporting the project history and its current development and exploitation state, the author analyses the required capability and quality attributes for a modern motion control system for robotics workshops and how GENERIS matches them. Thereafter, the GENERIS system capabilities and distributed architecture are described. The author concludes with the product organisation and the current and planned development activities.


Source Document: 1674 words  
Author Abstract: 49 words (2.92%)  
**SumUM**: 41 Words (2.45%)  

Enhanced servo motor technology through reduced size and weight and increased torque performance is now being incorporated within arc welding robots. Welding was implemented in the past by controlling the weld current during short-circuiting. Shows high speed arc welding technology and Servotorch.

**feed speed, heavy current, heavy welding current, high speed, high-speed arc, new arc, robots, stable arc, weld, welding current**

By implementing a high-speed welding system with a new arc welding robot, researchers have far exceeded current welding speeds. The heat warp of the work piece and adhesion of spatter onto the workpiece has been decreased, reducing the cycle time, and total heat input.

*Issues and experimental results in vision-guided robotic grasping of static or moving objects.*

Source Document: 3611 words
Author Abstract: 155 words (4.3%)
SumUM: 88 Words (2.43%)

Highlights the design choices the authors have made while developing a vision-guided grasping system, driven in part by the goal of high success rate grasping; presents the results from three sets of random object placement experiments that emphasize failure analysis in order to improve the systems performance and an overview of the work related to vision-guided grasping and the design decisions researchers made when addressing sensor system and robot system issues; and also reports the unresolved issues related to vision-based grasping in general and the system in particular. The desired control input was calculated.

64 experiments, MRVT system, basic design issues, canonical stereo system, control, experiments, good grasping system, grasping systems, Issues, monocular system, open-loop control, promising results, proposed systems, revised grasping system, robotic system, second set, stereo systems, system, vision system, vision-guided grasping system

Many research efforts have turned to sensing, and in particular computer vision, to create more flexible robotic systems. Computer vision is often required to provide data for the grasping of a target. Using a vision system for grasping of static or moving objects presents several issues with respect to sensing, control, and system configuration. This paper presents some of these issues in concept with the options available to the researcher and the trade-offs to be expected when integrating a vision system with a robotic system for the purpose of grasping objects. The paper includes a description of our experimental system and contains experimental results from a particular configuration that characterize the type and frequency of errors encountered while performing various vision-guided grasping tasks. These error classes and their frequency of occurrence lend insight into the problems encountered during visual grasping and into the possible solution of these problems.


Source Document: 3020 words
Author Abstract: 105 words (3.47%)
SumUM: 148 Words (4.90%)

The McKibben muscle was invented in the 1950s by the American atomic physicist Joseph L. McKibben with the aim of motorizing an arm orthotics for the poliomyelitic daughter. The forgotten invention was rediscovered in the 1980s by the Japanese pneumatic tyre manufacturer, Bridgestone, who had remarked the similarity between the artificial muscle components and pneumatic tyre ones. Since then other research teams have developed McKibben muscle
replicas which took different names: "Braided artificial muscles", "Digit muscle", "Pneumatic muscle". Describes the McKibben muscle and analyses the performances of robot-arms driven by the artificial muscles. A two d.o.f. robot actuated by McKibben muscles was designed to assess closed-loop control performances of the McKibben muscle actuator. A basic controller was designed to test the robot including two levels: a point-to-point trapezoidal velocity profile trajectory generator and a joint closed-loop control. Shows original use of McKibben muscle as forearm and force change principle.

**McKibben muscle, McKibben muscle replicas, high ratios maximum force, human arm, muscles, muscle maximum force, one adopted by Bridgestone robots, principle, robot, true artificial muscle, Typical McKibben muscle, use**

Describes the McKibben muscle and its major properties. Outlines the analogy between this artificial muscle and the skeletal muscle. Describes the actuator composed of two McKibben muscles set into antagonism based on the model of the biceps-triceps system, and explains its natural compliance in analogy with our joint liveness. Reports some control experiments developed on a two d.o.f. robot actuated by McKibben muscles which emphasize the ability of these robot-arms to move in contact with their environment as well as moving loads of high ratio to the robot's own weight. Also outlines control difficulties and accuracy limitations and discusses applications.


Source Document: 6229 words
Author Abstract: 134 words (2.15%)  
**SUM**: 160 Words (2.56%)

Takes a user centric approach to the design of the advertisement broker. The components were built using well understood technology: Web browser for a user interface, HTTP and RPC for communication between components, and a relational database as a repository for the catalog. Work was completed to satisfaction the authors can now look more deeply at the problem of advertising structured database services. A criticism of the resource discovery field is that it is driven by computing researchers who are focussed on narrow technical issues and tend to ignore existing knowledge about user search behavior. It demonstrates the baseline services a catalog should offer, and gives researchers an idea as to where novel extensions can be added. Covers descriptions in surrogates and development of resource discovery tools. Shows architecture.

**WWW interface, Advertisements, advertisement broker, catalog, database technology, effective surrogate, employee database, hand crafted surrogates, perfect surrogate, resources, Resource discovery research studies tools, Resource discovery tools, services, Surrogates, Systems researchers - technology, tools, users, user interface, Web, Web technology**
Discusses some of the problems designers face in building catalogs in large networks and relates them back to the resource discovery problem. Currently many catalogs tend to be built in an ad hoc fashion - which leads to a great variety in the quality of publicly accessible network catalogs. Furthermore, the research surrounding these catalogs tends to focus on narrow technical issues - resulting in difficult-to-use catalogs. Addresses this problem by providing a usability framework based on the library science and human computer interaction literature, and demonstrates some of those principles via an example of a prototype. Results are interesting to resource discovery tool developers in that a framework for understanding the general resource discovery problem is provided and some techniques for dealing with those problems are presented.


Source Document: 5194 words
Author Abstract: 81 words (1.55%)
SumUM: 88 Words (1.70%)

The National Aeronautics and Space Act of 1958 established the National Aeronautics and Space Administration and charged it to "provide for the widest practicable and appropriate dissemination of information concerning ... the activities and the results thereof". The goal of NTRS (Technical Report Server) is to provide "one-stop-shopping" for NASA (Space Administration) technical publications. The results of the other projects will be included in NTRS when they are available. Shows NTRS through NCSA (Supercomputing Applications) Mosaic for the X Window System and a formatted abstract.

NASA, NTRS, NTRS WWW page, NTRS home page, NTRS information, NTRS page, abstract, abstract server, casual computer users, field searches, freely available and highly popular NCSA Mosaic, information system, interface, most obvious is that NTRS, publication, systems, Technical Report Server, uniquely extensible Mosaic

The National Aeronautics and Space Act of 1958 established the National Aeronautics and Space Administration (NASA) and charged it to "provide for the widest practicable and appropriate dissemination of information concerning ... its activities and the results thereof". The search for innovative methods to distribute NASA’s information led a grassroots team to create the NASA Technical Report Server (NTRS), which uses the World Wide Web and other popular Internet-based information systems.


Source Document: 3430 words
Author Abstract: 109 words (3.17%)

SumUM: 48 Words (1.40%)

Presents the development of one such gripper, the 234-gripper. The goal of the project is to obtain a reliable, working demonstration-system of a flexible assembly cell that is capable of assembling industrial products. The research aims at the following aspects: The gripper will be integrated in the DIAC (Delft Intelligient Assembly Cell) assembly cell. Covers a description of the operations.

**DIAC, applicable grippers, Flexible grippers, generic industrial gripper, gripper, grippers, main controller grasps, stable grasp**

Describes current research work into the development of a three finger industrial gripper suitable for flexible assembly work. Outlines the key aspects of building and programming the gripper and ways of simplifying its control with local computing and looks at the interaction between product design, gripper properties and assembly process. Concludes that a generic industrial gripper has been built, that is as fast as a normal gripper and which is able to grasp more part shapes more stably. Programming of the gripper is partly automated to provide flexibility to the assembly process. Further research work is being carried out on the project.


Source Document: 4360 words

Author Abstract: 106 words (2.43%)

SumUM: 143 Words (3.28%)

For automation to be effective and enhance the competitiveness of a US shipyard the welding process requires sensor, processing of the sensor data, and adaptive control of the welding process. PAWS (the programmable automated welding system) was designed to provide an automated means of planning, controlling, and performing critical welding operations for improving productivity and quality. The system was designed specifically to support shipyard welding applications and the basic functionality in OLP can be customized to support specified models of PAWS. The motion for a feature was planned, the macro can be used to plan the motion for identical features which are in other positions on the part.

**B&W CIM Systems, OLP, VME-based robot controller, adept controller, controller, Conventional systems, conventional technology, feature, open architecture controller, operations, PAWS, PAWS controller, PAWS system, PAWS vision system, process, robots, robot motions, shipyards, simulation features, solid models, system, technologies, Total control, unique features, vision system, weld, welding process**
Reports on an aggressive project to develop an advanced, automated welding system, being completed at Babcock & Wilcox, CIM Systems. This system, the programmable automated welding system (PAWS), involves the integration of both planning and control technologies to address the needs of small batch robotic welding operations. PAWS is specifically designed to provide an automated means of planning, controlling, and evaluating critical welding situations in shipyard environments to improve productivity and quality. Five varieties (wall, lathe, floor mount, cantilevered, and gantry) of PAWS welding systems currently exist.


Source Document: 4624 words  
Author Abstract: 192 words (4.15%)  
**SumUM**: 82 Words (1.77%)

Presents the views on the development of intelligent interactive service robots. The authors have continually observed that a key research issue in service robotics is the integration of humans into the system. Describes HuDL (local autonomy) in greater detail; discusses system integration and the IMA (the intelligent machine architecture); and also gives an example implementation. Covers the evolution of a robot from a system with limited abilities and redundant implementation of common system elements. Shows technologies and basic behavior system for ISAC.

**HuDL, IMA, ISAC, aid systems, common system elements, development, elements, holonic manufacturing system, human, issue, key issue, local autonomy, robot, second issue, service robots, service robotics, system, Technologies**

For the past ten years, the Intelligent Robotics Laboratory (IRL) at Vanderbilt University has been developing service robots that interact naturally, closely and safely with human beings. Two main issues for research have arisen from this prior work. The first is how to achieve a high level of interaction between the human and robot. The result has been the philosophy of human directed local autonomy (HuDL), a guiding principle for research, design, and implementation of service robots. The human-robot relationship we seek to achieve is symbiotic in the sense that both the human and the robot work together to achieve goals, for example as aids to the elderly or disabled. The second issue is the general problem of system integration, with a specific focus on integrating humans into the service robotic system. This issue has led to the development of the Intelligent Machine Architecture (IMA), a novel software architecture specifically designed to simplify the integration of the many diverse algorithms, sensors, and actuators necessary for intelligent interactive service robots.


Source Document: 2225 words
Author Abstract: 199 words (8.94%)
SumUM: 71 Words (3.19%)

Previous equipment used on the slaughterboard and in the boning room has been primarily "hard" automation - specialised automated equipment that has been specifically designed to perform one task, within narrowly defined parameters. Describes the features of the meat processing robot, specific tasks the authors have automated on the lamb slaughtering line, in the lamb boning room, and in automating beef processing. All technology is patented nationally and internationally by Industrial Research Limited and Meat New Zealand. Shows the cut and the commercial robotic Y-cutting system.

Y-cutting, Y-cutting system, beef processing commercial systems, first task, Flexible automation, force feedback control system, Industrial Research, inverted dressing system, lamb, machine vision technology, meat processing, pelt removal task, robot, robotic primal handling system, room tasks, system, technologies, robots

Over the last five years we have successfully researched, designed, developed and commercialised the world’s first lamb and sheep dressing robots. Two have already been sold to commercial concerns. This has caused a paradigm shift in the way automation in meat processing can be viewed. In this paper we describe the lessons we have learned in robotic automation via projects in Y-cutting, ripdown, brisket clearing, opening cuts, handling of primal cuts and packing bagged meat pieces for lamb and sheep meat. All of these projects have been, or are about to be, trialed in operating plants processing export quality meat. These projects have involved the development of a programmable robot suitable for washdown environments, and of tooling to conduct specific dressing and handling tasks. Latest projects are applying this approach to automating certain beef processing tasks, and a beef processing robot has been constructed and is being installed for trials in an operating plant. The technology behind the robots is described and illustrated in our paper. Also described are the methods we used to ensure commercialisation was an economic success.


Source Document: 4978 words
Author Abstract: 115 words (2.31%)
SumUM: 169 Words (3.40%)

Taking into account the extreme complexity of the problems related to the environment and the limitations of the current approaches, the implementation of a fully automatic and real time solution for this task seems far away. Robotics approaches have been applied since the late 1970s with more and more advanced devices and strategies. The harvester has been tested in laboratory conditions : tests are described and results are given together with some conclusions of the work. Presents the mechanical and electronic design of the robot harvester including all subsystems, namely, fruit localisation module, harvesting arm and
gripper-cutter as well as the integration of subsystems. Strategies, such as centering the fruit in the image during the approximation movements toward the fruit, or stereoscopic vision with triangulation techniques from the matching of the images from two cameras for fruit location were implemented. Shows configuration of the robotic fruit harvester Agribot and schematic view of the detaching tool.

Agribot. Design, detaching cycle time, detaching operation, different modules, fruits, fruit localisation module, grasping operation, harvesting arm, modules, presented work. Robots, subsystems, work

This work presents a robot prototype designed and built for a new aided fruit-harvesting strategy in highly unstructured environments, involving human-machine task distribution. The operator drives the robotic harvester and performs the detection of fruits by means of a laser range-finder, the computer performs the precise location of the fruits, computes adequate picking sequences and controls the motion of all the mechanical components (picking arm and gripper-cutter). Throughout this work, the specific design of every module of the robotized fruit harvester is presented. The harvester has been built and laboratory tests with artificial trees were conducted to check range-finder’s localization accuracy and dependence on external conditions, harvesting arm’s velocity, positioning accuracy and repeatability; and gripper-cutter performance. Results show excellent range-finder and harvesting arm operation, while a bottleneck is detected in gripper-cutter performance. Some figures showing overall performance are given.


Source Document: 2696 words
Author Abstract: 25 words (1.00%)
SumUM: 56 Words (2.07%)

Presents the performance of 14 applications on the Alewife machine, including both coarse- and fine-grain applications; and also shows that Alewife provides excellent communication mechanisms for fine-grain applications, even without data reuse. Presents correlation of application rank for each metric relative to processor utilization; and also shows load balance for four applications and comparison of ICCG.

Alewife, Alewife machine, ICCG, applications, Application Performance, load balance, memory applications, memory machines, poor utilization, real scientific data, reasonable performance

We explore the performance of 14 applications on the Alewife machine, introducing two new performance metrics weighted cache-hit ratio and weighted computation granularity.

Facilitating designer-customer communication in the World Wide Web. T. Tuikka et al. In-
Virtual prototyping is a technique which has been suggested for use in, for example, telecommunication product development as a high-end technology to achieve a quick digital model that could be used in the same way as a real prototype. Presents the design rationale of WebShaman, starting from the concept design perspective by introducing a set of requirements to support communication via a concept model between industrial designer and a customer. In the paper, the authors suggest that virtual prototyping in collaborative use between designers is a potential technique to facilitate design and alleviate the problems created by geographical distance and complexities in work between different parties. The technique, which has been implemented in the VRP project, allows component level manipulation of a virtual prototype in a WWW (World Wide Web) browser. The user services, the software architecture, and the techniques of WebShaman have been developed iteratively during the fieldwork in order to illustrate the ideas and the feasibility of the system. The server is not much different from the other servers constructed to support synchronous collaboration.

3D model, Internet technology, VIRPI project, VRML 2.0 model, WWW, WWW techniques, WebShaman, CAD systems, concept, conceptual model, customers, designers, models, object-oriented model, product, product concept, product design, requirements, server, simulation model, Smart virtual prototypes, smart virtual prototyping techniques, software components, systems technique, technology, use, Virtual components, virtual prototype, virtual prototype interface model, virtual prototype system, virtual prototyping, work

Introduces a way to design geographically distributed virtual prototyping, a new Internet technology, in order to facilitate designer-customer communication in the product development of small electronic devices, such as mobile telephones. First, we will present our research in the concept design domain with a set of requirements focusing on communication between the designer and the customer. Second, a technique called "smart virtual prototyping" will be presented to elaborate on the virtual prototyping techniques to be used over the World Wide Web. Third, we will present the main ideas, architecture and selected software techniques of WebShaman, which is an application built to demonstrate how a distributed virtual prototyping system could support geographically distant designer-customer communication. Finally, we discuss the possible impact of the distributed virtual prototyping approach on the WWW community.


Source Document: 5392 words
Author Abstract: 118 words (2.18%)  
SumUM: 106 Words (1.96%)

Describes the overall architecture of WebACE, an agent for document categorization and exploration that operates on Web documents; and also explores the performance of the agent as an integrated and fully automated system, comparing the relative merits of various algorithms for clustering, query generation, and document filtering, when used as the key components for this agent. The main components of the agent were implemented by using a proxy server, enough to create a single-user client-side server prototype capable of clustering documents retrieved by the user. The authors then know that the new incoming document is most closely related to the paper already in that cluster. Presents implementation.

PDDP cluster, WebACE, WebACE agent, proxy server, agent, clusters, documents, leaf clusters, local proxy server, new documents, novel algorithms, components, proxy server, relevant documents

The authors propose a client-side agent for exploring and categorizing documents on the World Wide Web. As the user browses the Web using a usual Web browser, this agent is designed to aid the user by classifying the documents the user finds most interesting into clusters. The agent carries out the task completely automatically and autonomously, with as little user intervention as the user desires. The principal novel components in this agent that make it possible are a scalable hierarchical clustering algorithm and a taxonomic label generator. In this paper, the overall architecture of this agent is described and the details of the algorithms within its key components are discussed.