Approche hybride pour le résumé automatique de textes. Application à la langue arabe.

Mohamed Hedi Maaloul

Université Aix-Marseille
ÉCOLE DOCTORALE EN MATHÉMATIQUES
ET INFORMATIQUE DE MARSEILLE

Université de Sfax
FACULTÉ DES SCIENCES ÉCONOMIQUES
ET DE GESTION DE SFAX

T H È S E

pour obtenir le titre de

Docteur en Informatique

Présentée par

Mohamed Hédi MAÂLOUL

Approche hybride pour le résumé automatique de textes. Application à la langue arabe.

soutenue le 18 décembre 2012.

Jury :

Président :	Alexis NASR	LIF (Université Aix-Marseille)
Rapporteurs :	Rim FAIZ	IHEC (Université de Carthage)
	Juan Manuel TORRES-MORENO	LIA (Université d'Avignon)
Directeurs :	Lamia HADRICH BELGUITH	FSEGS (Université de Sfax)
	Philippe BLACHE	CNRS - LPL (Université Aix-Marseille)

Année Universitaire : 2011-2012

* وَ فَوْقَ كُلِّ ذِي عِلْمٍ عَلِيمٌ *

صَدق اللّهُ العظيم

﴿سورة يوسف آية ٧٦﴾

* *Et au-dessus de tout homme détenant la science il y a un savant* *

﴿ Joseph 76 ﴾

Dédicaces

Mes dédicaces ne sont que l'expression de mes profondes gratitudes, de mes salutations chaleureuses et de ma sincère reconnaissance à tous ceux qui comblent ma vie et y confèrent son goût et sa saveur.

Je dédie cette thèse à :

Mon père

Ce symbole de sacrifice et de dévouement, ses aides et ses recommandations m'ont souvent incité à persévérer dans l'effort et à progresser dans ma vie universitaire et professionnelle.

Ma mère

Ce rayon de soleil qui ne cesse d'éclairer ma vie, cette source inépuisable d'amour, de tendresse et d'affection, grâce à ses conseils, j'ai pu frayer mon chemin et savourer le goût de la réussite et du succès.

Ma fiancée Rania

Qui m'a soutenu moralement et sentimentalement. Elle constitue mon seul et unique soulagement, ses paroles me redonnent souvent confiance en mes aptitudes, m'exhortent à exceller et à défier toutes les entraves et me comblent d'optimisme et d'espoir.

Mes sœurs et leurs conjoints Abderrazek et Ahmed

Pour leur précieux encouragement et leur continuelle stimulation.

Mon beau-père Hédi et ma belle-mère Monia

Pour leur précieux encouragement et soutien.

Mes proches et surtout mon oncle Abdessattar et mes amis
Raouf, Najah[2], Abdallah, Omar, Hatem, Tijani, Iheb, Nabil et Jamil.

Chez qui j'ai trouvé le sein toujours prêt à me consoler et le cœur bienveillant apte à mettre fin à mes maux et soucis.

À toute ma famille

À tous ceux qui me sont chers

Remerciements

En rédigeant cette page du manuscrit, je suis obligé de reconnaître que cette thèse est le fruit d'un peu de recherche et de beaucoup d'aide reçue de nombreuses personnes que je tiens à remercier.

Je tiens tout d'abord à remercier mes directeurs de thèse, Madame **Lamia Hadrich Belguith** Maître de conférences habilitée à diriger des recherches à la Faculté des Sciences Économiques et de Gestion de Sfax (FSEGS) et directeur du groupe de recherche ANLP [1] (Arabic Natural Language Processing Group) du laboratoire MIRACL [2] (Multimedia, InfoRmation systems and Advanced Computing Laboratory) et Monsieur **Philippe Blache** Directeur de Recherche au CNRS, affecté au LPL [3] (Laboratoire Parole et Langage), qui ont bien voulu m'aider et m'encadrer tout au long de la réalisation de ce travail malgré leurs tâches très lourdes. Je tiens à les remercier vivement pour les conseils fructueux qu'ils m'ont prodigués, pour leurs orientations ciblées, pour leur disponibilité, même à des heures peu adaptées, et pour la confiance qu'ils ont su m'accorder.

Merci également aux membres de mon jury, pour avoir accepté de participer à ce jury et pour l'intérêt qu'ils ont exprimé pour ce travail. Je suis très honoré et je suis très reconnaissant à Monsieur **Alexis Nasr**, professeur au Laboratoire d'Informatique Fondamentale (LIF) et président du jury ; aux deux rapporteurs de ce travail, Madame **Rim Faiz** professeur à l'Institut des Hautes Études Commerciales de Carthage (IHEC de Carthage) et Monsieur **Juan Manuel Torres-Moreno** Maître de Conférences Habilité à diriger des recherches au Laboratoire Informatique d'Avignon (LIA).

Parmi les nombreux permanents du LPL, je tiens à remercier tout particulièrement, les ex-membres du bureau des doctorants (Céline De Looze, Amina Chentir et Vincent Aubanel) pour leur bonne humeur et leurs conseils toujours avisés.

Puis viennent les membres du groupe ANLP, toujours aussi nombreux. Ainsi, je tiens à remercier Younes Bahou, Docteur Maher Jaoua, Docteur Hatem Hadj kacem, Docteur Chafik Aloulou et Docteur Mariem Ellouze pour leurs conseils toujours pointus et avisés.

1. https://sites.google.com/site/anlprg/
2. http://www.miracl.rnu.tn/
3. http://www.lpl.univ-aix.fr/

Je tiens évidemment à remercier ma famille et mes amis, qui tout au long de ces 1735 jours de thèse m'ont aidé à garder le cap et à croire que tout cela aurait une fin. Mille fois merci !

Résumé

Face à l'avènement de l'Internet et des moteurs de recherche, la quantité de textes disponibles en format électronique est devenue énorme. Il est donc indispensable d'offrir des outils de visualisation rapide des textes, en particulier des résumés automatiques, afin que l'utilisateur puisse évaluer la pertinence d'un document vis-à-vis de l'information recherchée.
Un résumé est un texte concis qui rend compte du contenu "essentiel" d'un autre texte, dit texte source [Saggion 2000]. Le but du résumé est d'aider le lecteur à décider si le document source contient l'information recherchée ou pas. Réciproquement, si l'information recherchée existe dans le résumé, le lecteur peut ne pas avoir besoin de lire le document entier.
Plusieurs approches ont été explorées pour le résumé automatique, les unes basées sur des techniques linguistiques (basées sur l'analyse du discours et de sa structure), les autres sur des approches statistiques (basées sur la distribution des occurrences des mots) [Amini 2003]. Cependant, la plupart des travaux dans le domaine du résumé sont basés sur l'extraction, même si la lecture des résumés par extraction peut paraître difficile en raison du manque de cohérence.
Cette thèse s'inscrit donc dans le cadre du Traitement Automatique du Langage Naturel (TALN) et plus précisément celui du résumé automatique de documents arabes. Nous avons fixé comme objectif le développement de techniques *hybrides* (*symboliques*/*numériques*) pour le résumé automatique de documents. Nous appliquons ces techniques à la langue arabe.
La première phase de notre travail a été principalement consacrée à l'étude de l'état de l'art dans le domaine du résumé automatique et plus précisément des méthodes d'extraction fondées sur les techniques *symboliques* et/ou *numériques*.
La deuxième phase de notre travail s'est intéressée à la proposition d'une approche *hybride* faisant appel à des techniques *symboliques*/*numériques* pour le résumé automatique de documents. Ainsi, pour sélectionner les phrases du résumé, l'approche à proposer consiste à faire interagir :

(1) des techniques symboliques d'analyse du discours et de sa structure qui se basent sur une analyse rhétorique. Celle-ci a pour rôle de mettre en évidence dans un texte source les unités minimales (noyaux – segments de texte primordiaux pour la cohérence et qui charpentent un discours) nécessaires pour le processus d'extraction,

(2) des techniques numériques qui se reposent sur un mécanisme d'apprentissage basé sur l'algorithme SVM (Support Vector Machine). Cet apprentissage permet de déterminer parmi les phrases non porteuses d'information rhétorique et ayant des relations "Autres - آخر" (i.e. la relation rhétorique "Autres - آخر" est attribuée lorsqu'aucune relation rhétorique n'est déterminée) celles qui sont pertinentes pour l'extrait final. Remarquons que les critères numériques ont l'avantage d'être portables, applicables à tout type de corpus, et ont

déjà fait leurs preuves pour d'autres langues comme l'anglais et le français [Lehmam 1995].

Soulignons que cette approche aborde la question des besoins des utilisateurs : une information n'est en effet pas importante en soi, mais doit correspondre aux besoins d'un utilisateur. Notre recherche s'oriente en particulier vers la production de résumés dynamiques [Maaloul 2010a].

Nous avons choisi l'arabe comme langue d'application. Ce choix se justifie d'une part par la rareté des travaux similaires dans ce domaine et d'autre part, par les objectifs fixés au sein de notre équipe de recherche ANLP [4] : contribution aux recherches sur le traitement de la langue arabe, en mettant à la disposition des chercheurs des outils d'analyse et de traitement automatique de l'arabe.

Afin de montrer la faisabilité de notre approche, nous avons développé le système "L.A.E (Lakas Al El'eli) - الـلَخَاص الآلي". Notre implantation se base, en partie, sur l'utilisation d'un certain nombre d'outils (le segmenteur "STAr"et l'analyseur morphologique "MORPH- 2").

L'évaluation de notre système "L.A.E (Lakas Al El'eli) - الـلَخَاص الآلي" a porté sur un corpus de référence constitué de 150 articles de presse. Nous disposons pour ce corpus du jugement de deux experts (accord inter-annotateurs de 0,68). L'évaluation a donné des valeurs de 0.47 de rappel, 0.61 de précision et 0.53 de F-mesure.

Un des points forts de l'approche proposée réside dans deux aspects :

- interaction de connaissances purement linguistiques avec d'autres purement numériques,
- traitement d'articles de presse portant sur des thèmes hétérogènes.

L'utilisation du mécanisme d'apprentissage a permis de remédier au problème de manque d'information linguistique. Il a ainsi résolu le problème d'une phrase pertinente non retenue dans l'extrait final car non porteuse de relation rhétorique.

4. http://sites.google.com/site/anlprg

Table des matières

I État de l'art **5**

1 L'activité résumante **7**

1.1 Introduction . . . 7
1.2 Qu'est ce qu'un résumé ? . . . 8
1.3 Processus humain pour le résumé . . . 9
1.3.1 Description du processus de résumé par compréhension . . . 9
1.3.2 Facteurs influençant la forme du résumé . . . 11
1.4 Différents types du résumé textuel . . . 11
1.5 Types d'utilisateurs des résumés . . . 13
1.6 Caractéristiques des résumés . . . 13
1.6.1 Concision . . . 14
1.6.2 Couverture . . . 14
1.6.3 Fidélité . . . 14
1.6.4 Cohésion et cohérence . . . 14
1.7 Motivations pour l'automatisation du résumé . . . 15
1.8 Conclusion . . . 15

2 État de l'art sur les approches de résumés automatiques **17**

2.1 Introduction . . . 18
2.2 Approche numérique . . . 18
2.2.1 Méthodes axées sur les calculs statistiques . . . 19
2.2.2 Méthodes fondées sur l'apprentissage . . . 22
2.2.3 Discussion des méthodes numériques . . . 29
2.3 Approche symbolique . . . 29
2.3.1 La grammaire de Montague . . . 30
2.3.2 La théorie de la structure rhétorique . . . 31
2.3.3 La théorie de la représentation discursive . . . 34
2.3.4 Techniques de la théorie de la représentation discursive segmentée (SDRT) 36
2.3.5 Techniques de l'exploration contextuelle . . . 40
2.3.6 Discussion des méthodes symboliques . . . 41
2.4 Approche hybride . . . 43
2.5 Conclusion . . . 43

II Les bases théoriques et techniques pour une nouvelle approche 45

3 Les techniques de TALN pour le résumé automatique de l'arabe 47
3.1 Introduction . . . 47
3.2 Particularités de la langue arabe . . . 48
3.2.1 Absence de voyelles . . . 49
3.2.2 Agglutination . . . 50
3.2.3 Irrégularité de l'ordre des mots dans la phrase . . . 50
3.2.4 Absence de ponctuation régulière . . . 51
3.3 Difficultés de l'analyse automatique de l'arabe . . . 51
3.3.1 La segmentation de textes . . . 51
3.3.2 L'analyse morphologique . . . 51
3.3.3 L'étiquetage grammatical . . . 52
3.3.4 L'analyse syntaxique . . . 53
3.4 Principales approches de traitement automatique de l'arabe écrit . . . 54
3.4.1 Approches de Segmentation de textes . . . 54
3.4.2 Approches d'analyse syntaxique . . . 56
3.4.3 Approches d'étiquetage grammatical . . . 58
3.4.4 Approches d'analyse morphologique . . . 58
3.4.5 Approches de reconnaissance des entités nommées . . . 59
3.5 Conclusion . . . 61

4 Annotation rhétorique 63
4.1 Introduction . . . 63
4.2 Gestion des connaissances linguistiques . . . 64
4.2.1 Construction du corpus . . . 64
4.2.2 Étude du corpus . . . 65
4.2.3 Organisation des frames rhétoriques en relations . . . 71
4.2.4 Règles de correction des relations rhétoriques . . . 72
4.3 Étapes de la méthode d'annotation rhétorique . . . 73
4.3.1 Détermination de la relation rhétorique et de la nature de l'unité minimale 75
4.3.2 Enrichissement des relations rhétoriques . . . 75
4.3.3 Correction des relations rhétoriques . . . 75
4.3.4 Détermination de l'arbre RST le plus descriptif . . . 76
4.4 Conclusion . . . 78

5 Génération d'extrait par une approche hybride 79
5.1 Introduction . . . 79

5.2 Notre proposition : étapes de l'approche proposée . . . 80
5.2.1 Segmentation du document source . . . 81
5.2.2 Étiquetage morphologique . . . 82
5.2.3 Analyse rhétorique . . . 83
5.2.4 Phase d'apprentissage . . . 84
5.2.5 Sélection et classement des phrases selon le type du résumé . . . 90
5.3 Conclusion . . . 92

III Un système hybride de résumé automatique 93

6 Architecture et description du système 95
6.1 Introduction . . . 95
6.2 Architecture du système . . . 96
6.3 Implémentation : du texte brut à l'extrait . . . 97
6.3.1 Chargement du texte . . . 98
6.3.2 Segmentation du texte . . . 99
6.3.3 Étiquetage morphologique . . . 100
6.3.4 Analyse rhétorique du texte . . . 102
6.3.5 Détermination de l'arbre RST . . . 103
6.3.6 Sélection des critères d'extraction reflétant les besoins de l'utilisateur . . . 104
6.3.7 Sorties du système . . . 106
6.4 Conclusion . . . 107

7 Expérimentation et validation 109
7.1 Introduction . . . 109
7.2 Les métriques d'évaluation utilisées dans les compagnes DUC/TAC . . . 110
7.2.1 Les mesures ROUGE . . . 112
7.2.2 Les mesures PYRAMID . . . 112
7.2.3 Les mesures de rappel et de précision . . . 113
7.2.4 Les mesures d'évaluation linguistique . . . 114
7.3 Évaluation de l'outil d'annotation rhétorique . . . 114
7.3.1 Protocole expérimental . . . 115
7.3.2 Apport de l'étiquetage morphologique dans l'annotation rhétorique . . . 117
7.4 Évaluation du processus de génération d'extrait . . . 118
7.5 Conclusion . . . 119

Bibliographie 127

A Relations rhétoriques **143**

B Critères d'extraction **169**

Table des figures

1.1 Processus humain pour le résumé par compréhension [Giquel 1990] 10

2.1 Principe des systèmes à apprentissage supervisé [Minel 2002b]. 24
2.2 Principe des systèmes à apprentissage semi-supervisé [Amini 2001] 27
2.3 Cinq modèles de schémas rhétoriques [Mann 1988] 32
2.4 Arbres RST de l'exemple 1 33
2.5 Arbres RST de l'exemple 2 34
2.6 La représentation de la DRS_K 35
2.7 Représentation de SDRS hiérarchisée [Busquets 2001] 38
2.8 Représentation et graphe de SDRS [Busquets 2001] 39

4.1 Phases d'analyse rhétorique 74
4.2 Arbre RST 78

5.1 Étapes de l'approche hybride proposée [Maaloul 2011] 81
5.2 Principe d'apprentissage 85
5.3 Exemple de texte étiqueté : texte source avec son résumé de référence 86
5.4 Support Vector Machine 87

6.1 Architecture du système "L.A.E - الـلَخَاص الآلي" 96
6.2 Interface principale du système "L.A.E - الـلَخَاص الآلي" 98
6.3 Interface de chargement du texte 99
6.4 Segmentation du texte en titres, sous-titres et phrases. 100
6.5 Exemple de fichier XML délivré par le module de segmentation 101
6.6 Étiquetage morphologique du texte 102
6.7 Analyse rhétorique du texte 103
6.8 Interface de création de l'arbre RST le plus descriptif 104
6.9 Interface utilisateur pour la sélection des critères d'extraction 105
6.10 Interface de génération de l'extrait final 106
6.11 Extrait final en format XML 107

7.1 Paramètres de calcul de la précision et du rappel [Minel 2002a] 113

Liste des tableaux

2.1 Exemple de critères d'extraction utilisés pour l'apprentissage [Mani 1998] 25
2.2 Exemples de relations rhétoriques et de marques associées [Ono 1996] 32
2.3 Exemple d'énoncé avec sa représentation formelle 35

3.1 Exemple de voyellation [Debili 2002] 49
3.2 Exemple de combinaisons possibles d'inversion de l'ordre des mots dans la phrase [Belguith 2005] 50
3.3 Exemple d'étiquettes grammaticales attribuées selon la voyellation [Debili 2002] 53
3.4 Exemple d'une règle de segmentation relative à la virgule 55
3.5 Corpus d'évaluation de STAr [Belguith 2005] 55
3.6 Résultat de l'évaluation de STAr [Belguith 2005] 56

4.1 Liste des relations rhétoriques définies par Asher et Lascarides pour un discours narratif [Asher 2003] 67
4.2 Liste des relations rhétoriques 67
4.3 Exemple de frame utilisé pour la détection de la relation rhétorique "Spécification - تخصيص" 68
4.4 Exemple de frame utilisé dans l'exemple 1, pour la détection de la relation rhétorique "Pondération - ترجيح" 70
4.5 Exemple de frame utilisé dans l'exemple 2, pour la détection de la relation rhétorique "Affirmation - جزم" 70
4.6 Exemple de frame utilisé pour la détection de la relation rhétorique "Négation - نفي" 71
4.7 Exemple de frame utilisé pour la détection de la relation rhétorique "Confirmation - توكيد" 72
4.8 Exemple de frame utilisé pour la détection de la relation rhétorique "Pondération - ترجيح" 72
4.9 Exemple de frame utilisé pour la détection de la relation rhétorique "Évidence - قَاعدة" 77
4.10 Exemple de frame utilisé pour la détection de la relation rhétorique "Condition - شرط" 77

5.1 Liste des balises HTML de mise en forme utilisées pour la détection des passages début et fin 82
5.2 Caractéristiques du corpus d'apprentissage 86

5.3 Liste des relations rhétoriques retenues pour le type de résumé indicatif, informatif et opinion 91

6.1 Critères de choix 107

7.1 Les tâches d'évaluation proposées dans les conférences DUC et TAC [Torres-Moreno 2011] 111
7.2 Degré d'accord et valeur de Kappa proposés par Landis et Koch [Landis 1977] . 116
7.3 Résultats des performances de l'outil d'annotation rhétorique 117
7.4 Moyennes des mesures de Rappel, Précision et F-mesure 119

Introduction générale

Face à l'avènement de l'Internet et des moteurs de recherche, l'information textuelle en format électronique s'accumule rapidement et en très grande quantité. De ce fait, il est intéressant d'offrir des outils informatiques de visualisation rapide des textes, comme par exemple des résumés automatiques (en condensant les textes de façon pertinente) afin que l'utilisateur puisse évaluer la pertinence d'un document vis-à-vis de l'information recherchée.

Un résumé est un texte concis qui rend compte du contenu "essentiel" d'un autre texte, dit texte source [Saggion 2000]. En effet, le but du résumé est d'aider le lecteur à décider si le document source contient l'information recherchée ou pas. Il se peut aussi que le lecteur n'ait pas besoin de lire la totalité du document source simplement parce que l'information recherchée existe dans le résumé.

Remarquons que le résumé automatique a inspiré de diverses orientations. En effet, plusieurs approches ont été explorées en linguistique (basées sur l'analyse du discours et de sa structure) [Mathkour 2008] et en statistique (basées sur la distribution des occurrences des mots) [Amini 2003]. Cependant, la plupart des travaux dans le domaine du résumé sont basés sur l'extraction, même si la lecture des résumés par extraction puisse être difficile en raison du manque de cohérence [Douzidia 2004a].

Ainsi, c'est dans le cadre du Traitement Automatique du Langage Naturel (TALN) et plus précisément celui du résumé automatique de documents arabes que s'inscrit ce sujet de thèse. Nous nous fixons comme objectif d'étudier et de proposer une approche hybride (symbolique/numérique) pour le résumé automatique de documents. Nous appliquons cette approche à la langue arabe.

De nos jours, la plupart des systèmes de résumé automatique traitent des textes en langues indo-européennes (l'anglais, le français, etc.). Le besoin de développer des systèmes de résumé automatique dédiés pour la langue arabe devient de plus en plus incontournable ces dernières années vu l'augmentation du nombre de documents électroniques rédigés en arabe.

En effet, nous avons recensés uniquement cinq travaux : le système Lakhas [Douzidia 2004a], le système EXCOM [Alrahabi 2006a], Arabic text summarization [Mathkour 2008, Azmi 2012], Arabic Query-Based Text Summarisation System (AQBTSS) et Arabic Concept-Based Text Summarisation System (ACBTSS) [El-Haj 2011].

Par ailleurs, il existe quelques outils industriels de résumé pour la langue arabe ou adaptables à l'arabe (Siraj [5], Essential Summarizer [6]).

Les méthodes proposées pour ces outils de résumé automatique se basent sur une approche

5. http://textmining.sakhr.com/
6. https://essential-mining.com

numérique ou symbolique. Ainsi par exemple l'outil Lakhas repose sur une méthode purement statistique qui utilise différents critères pour calculer les poids des phrases du texte source afin de permettre la sélection des phrases les plus pesantes dans le résumé [Douzidia 2004b]. Le système EXCOM, de filtrage sémantique de textes arabes, est basé sur la méthode de l'exploration contextuelle qui s'appuie sur des connaissances linguistiques et permet de repérer, grâce à des indices linguistiques, des informations pertinentes, comme les annonces thématiques, les énoncés définitoires, les titres, les soulignements, les récapitulations, etc. [Alrahabi 2006b].
Dans une lignée visant l'amélioration des résumés de type "extrait", nous proposons une méthode qui utilise une approche hybride faisant intervenir des techniques symboliques/numériques pour le résumé automatique de documents. Le genre textuel, auquel nous nous sommes intéressés, est les articles de presse.
Notons que l'expression "génération de résumé" n'évoque pas le processus classique préconisé par les applications de génération de langage. Dans le cadre de cette thèse, la génération de résumés se fait par interaction entre : i) des techniques symboliques d'analyse du discours et de sa structure, qui se basent généralement sur une représentation formelle des connaissances contenues dans le document source et ii) des techniques numériques qui se basent sur les statistiques, les probabilités et l'apprentissage. Signalons que les critères numériques ont l'avantage d'être portables, applicables à tout type de corpus, et ont déjà fait leurs preuves pour d'autres langues comme l'anglais [Edmundson 1969] et le français [Lehmam 1995].
Notre approche se base alors sur une analyse rhétorique du texte que nous plaçons avant le processus de sélection des éléments dits importants. Cette analyse donne une vision linguistique de la structuration des textes afin de détecter les relations sémantiques et les relations intentionnelles qui existent entre les segments textuels. En effet, cette analyse rhétorique a comme but d'établir les relations et les dépendances ainsi que l'importance relative des phrases ou propositions les unes par rapport aux autres [Teufel 1998]. Pour décrire notre approche, nous pouvons dire que les résumés automatiques seront générés par instanciation d'un type de résumé, choisi par le lecteur.
Ce rapport est organisé en de trois parties. Dans la première partie nous tentons de cerner l'objet de notre étude, à savoir le résumé et l'éventail d'approches existantes pour sa génération d'une façon automatique. Nous introduisons alors la terminologie nécessaire et nous examinons le processus de production des résumés chez les humains. Les résumés produits doivent être adaptés à la situation et aux besoins des utilisateurs. Nous décrivons alors les différents types de résumés générés selon ces paramètres. Par la suite, nous donnons une illustration assez exhaustive des approches existantes pour la génération d'un extrait. Cette présentation met en relief deux grandes familles d'approches à savoir l'approche numérique et l'approche symbolique. Enfin, nous concluons cette première partie par une discussion.
Nous décrirons dans la seconde partie les techniques de TALN pour le résumé automatique de

l'arabe. Pour cela, nous commençons le troisième chapitre par une présentation des particularités de la langue arabe et des principales ambiguïtés rencontrées lors de son analyse. Ensuite, nous donnons un aperçu sur les approches d'analyse de l'arabe se rapportant aux niveaux lexical, syntaxique et sémantique.
L'annotation rhétorique de textes sera décrite en détails dans le quatrième chapitre avec une présentation préalable et rapide sur l'intérêt de l'annotation dans le résumé automatique d'un article de presse. Nous en viendrons alors à présenter les différentes phases de notre proposition pour l'annotation rhétorique qui repose essentiellement sur la théorie de la structure rhétorique – RST (Rhetorical Structure Theory). Ensuite, nous décrirons l'avantage de l'utilisation des frames rhétoriques, morphologiques et des règles de correction dans la détermination d'un seul arbre RST jugé le plus descriptif de l'organisation structurelle du texte source.
Après avoir donné des détails sur l'annotation rhétorique d'un document qui constitue la base essentielle de notre travail, son insertion dans une approche hybride de résumé automatique sera alors présentée dans le chapitre cinq. Nous donnerons une description du processus de génération d'extrait par une approche hybride, en décomposant les différentes étapes qui le forment. Nous nous attarderons sur les caractéristiques les plus importantes de notre travail.
Dans la troisième partie, des descriptions plus techniques seront faites dans le chapitre six, pour les différents traitements prenant place dans le processus résumant général L.A.E. Ensuite, deux évaluations portant sur l'analyse rhétorique et sur la qualité des extraits, que notre approche produira, seront présentées dans le chapitre sept. Ces deux évaluations consisteront, à partir des métriques d'évaluation, à comparer nos relations rhétoriques détectées et nos extraits face à d'autres élaborés par deux experts humains. À la suite de l'évaluation et de l'analyse des résultats obtenus, nous insisterons encore une fois sur les points les plus importants de notre approche afin de justifier ses atouts particuliers et aussi complémentaires par rapport à d'autres approches numériques et symboliques.
Enfin, nous conclurons cette thèse en réaffirmant l'intérêt de notre approche hybride dans le cadre général de génération d'extrait selon les besoins d'un utilisateur. Nous ferons la synthèse des principaux points qui caractérisent notre travail. Comme perspectives de nos travaux nous envisageons d'étudier la possibilité d'adapter notre approche à d'autres langues.

Première partie

État de l'art

Chapitre 1

L'activité résumante

Sommaire

1.1 Introduction 7
1.2 Qu'est ce qu'un résumé ? 8
1.3 Processus humain pour le résumé 9
1.3.1 Description du processus de résumé par compréhension 9
1.3.2 Facteurs influençant la forme du résumé 11
1.4 Différents types du résumé textuel 11
1.5 Types d'utilisateurs des résumés 13
1.6 Caractéristiques des résumés 13
1.6.1 Concision 14
1.6.2 Couverture 14
1.6.3 Fidélité 14
1.6.4 Cohésion et cohérence 14
1.7 Motivations pour l'automatisation du résumé 15
1.8 Conclusion 15

1.1 Introduction

Dès la fin des années 1950 [Luhn 1958], est apparu le premier besoin de vouloir résumer automatiquement des documents. Ainsi, le besoin d'un logiciel de résumé automatique s'est progressivement fait ressentir suite à des besoins réels quant à la gestion de grosses masses textuelles sous formats numériques.

En effet, l'idée d'affecter la tâche de résumer des documents à une machine plutôt qu'à des humains est alors devenue à l'heure actuelle de plus en plus pressante sous l'effet de la nécessité de consulter rapidement une grande masse croissante de documents (textes techniques et scientifiques, courriers électroniques, articles de presse, etc.) ajoutée à cela une demande encore plus forte de la part des chercheurs universitaires, ingénieurs industriels, etc. [Blais 2008].

Entamer la lecture de cette grande quantité de documents pour chercher une information pertinente est une tâche très ardue, voire inconcevable. Le résumé automatique permet de faciliter

énormément cette tâche. En effet, le résumé synthétise les idées clés et importantes contenues dans le document. À l'issu de ces idées clés jugées importantes, le lecteur peut retenir ou rejeter le document.
Dans cette partie, nous tentons de cerner l'objet de notre étude, à savoir le résumé des textes. Nous débutons ainsi cette première partie par quelques définitions de cet objet. Nous étudierons par la suite l'activité humaine concernant le résumé. Cette activité résumante peut donner lieu à des produits de natures différentes selon un certain nombre de paramètres et de critères. Nous exposons ensuite les différents types de résumés et les caractéristiques qui leur sont inhérentes. Il est à noter que les éléments de cette partie se rapportent aussi bien au résumé produit par un agent humain que par une machine.

1.2 Qu'est ce qu'un résumé ?

Un résumé d'après le dictionnaire Le petit Larrouse [Collectif 2009] est une "Forme condensée d'un texte, d'un discours, etc. ; abrégé, sommaire.". Selon [Mani 1999], résumer consiste à condenser l'information la plus importante provenant d'un document (ou de plusieurs documents) afin d'en produire une version abrégée pour un utilisateur (ou plusieurs utilisateurs) et une tâche (ou plusieurs tâches). De même un résumé de texte peut être défini en tant qu'objet dont la taille est inférieure au texte source et dans lequel on retrouve présentes certaines des idées essentielles du texte d'origine [Masson 1998].
D'une façon générale, résumer quelque chose de la part d'un sujet, c'est en donner une représentation plus réduite et condensée, tout en en conservant l'essentiel. Ainsi, l'objet *résumé* est une substance textuelle qui peut être formée d'une entité distincte de sa source comme elle peut être une paraphrase assez courte de ce qu'a été mentionné dans le texte source. Ces deux formes de textualisation d'un résumé ont été évoquées par plusieurs travaux sur le résumé. Par exemple, dans [Hahn 1998b] et dans [Mani 2001b] on trouve des définitions terminologiques de ces deux formes. Ainsi, un résumé est dit *extrait* s'il consiste entièrement en matière linguistique copiée à partir du texte source. Il est désigné *abstract* s'il s'agit d'un texte en langue naturelle réécrivant le contenu du texte source en un nouveau texte plus compact tout en retenant les fragments d'information les plus pertinents [Mani 2001a].
Par ailleurs, Hahn [Hahn 1998a] présente la notion de *condensé* comme troisième forme de résumé. Le condensé n'est pas à vrai dire une forme textuelle concrète. C'est plutôt une structure de représentation formelle à partir de laquelle des « abstracts » peuvent être dérivés.
L'activité résumante relève donc d'une modification d'un texte source pour donner lieu à un nouveau texte, le texte cible [ROUX 1993]. Le terme "activité résumante" désigne une démarche cognitive complexe qui suppose une sélection et une hiérarchisation des informations contenues dans le texte source et facultativement une paraphrase résumante, c'est-à-dire, faisant appel à

des procédés de réduction des segments du texte retenus comme essentiels. Dans ce qui suit, nous allons analyser cette activité.

1.3 Processus humain pour le résumé

Élaborer un résumé est une tâche complexe. Elle nécessite une analyse en profondeur du contenu textuel et des capacités intellectuelles afin de dissocier ce qui est facultatif de ce qui est essentiel. En effet, la construction d'un résumé résulte d'une sélection précise d'informations du texte source sur des critères particuliers (suite à une compréhension), et elle ne procède pas par la sélection de certains éléments du texte source sur des critères mathématiques (statistiques), ou aléatoires. Toutefois, nous montrerons également qu'il existe des techniques et des méthodes numériques permettant de résumer sans analyse en profondeur du contenu, c'est-à-dire sans compréhension totale du texte source. Ces techniques et méthodes numériques, qu'emploient les résumeurs professionnels, ont aussi influencé certaines approches dans le domaine du résumé automatique[1]. Ces cas particuliers d'applications sont motivés par des raisons de vitesse, de productivité, et de souplesse des traitements qui sont engagées dans la construction de résumés. Néanmoins, dans l'activité résumante que l'homme effectue très régulièrement sur tout type d'objet (textes, films, évènements,etc.), la compréhension est nécessaire et est partie intégrante de l'élaboration d'un résumé [Blais 2008].
Dans ce qui suit, nous allons exposer le processus entrepris par un humain pour générer un résumé. La forme du résumé produit peut être influencée par plusieurs facteurs et critères particuliers. Nous présentons ces facteurs et critères vers la fin de cette section.

1.3.1 Description du processus de résumé par compréhension

Selon Giquel [Giquel 1990], l'élaboration du résumé par compréhension (voir figure 1.1) se compose de cinq étapes. Ces étapes sont : la lecture du texte source, l'analyse de l'information, la hiérarchisation de l'information, la synthèse de l'information et la rédaction du texte de résumé. Dans ce qui suit, nous allons détailler ces étapes.

- **Lecture complète du texte source**
  L'auteur du résumé est invité à la lecture complète et totale du texte une ou plusieurs fois pour en prendre connaissance. Il s'assure de la compréhension du vocabulaire. En lisant le texte, il se pose la question du contenu du texte, de son sens. Il observe la mise en pages du texte : ses titres, ses paragraphes, ses alinéas, etc. Après cette lecture, Il s'attarde plutôt sur certaines parties (début et fin en général), et sa lecture du texte n'est pas toujours linéaire [Giquel 1990].

1. Cf. partie suivante.

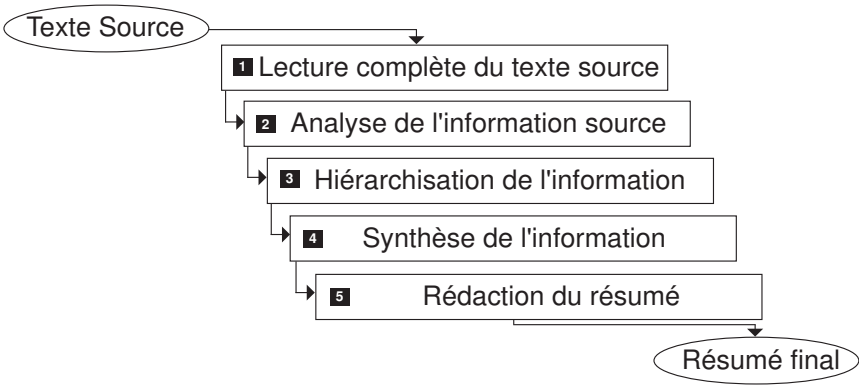

FIGURE 1.1 – Processus humain pour le résumé par compréhension [Giquel 1990]

– **Analyse de l'information source**
  Suite à la lecture complète du texte source, le lecteur réalise une analyse de l'information (analyse de certains passages, des mots-clés, des mots de liaison, des connecteurs, etc.). Le lecteur dresse l'inventaire de toutes les informations et rien que les informations contenues dans le texte-source. Il se pose la question du sens de chaque notion développée, en restant neutre et objectif. Il peut marquer les grandes parties du texte ou donner un titre à chacune de ces parties. Grâce à ces connaissances, le lecteur sait distinguer les éléments informatifs les plus importants qui constitueront le résumé [van Dijk 1983].
– **Hiérarchisation de l'information**
  La compréhension d'un texte par un sujet humain commence, dans un premier temps, par une opération de tri des informations présentées. En effet, l'auteur classe les idées fondamentales pour la bonne compréhension du texte et dégage les liaisons sémantiques exprimant la cohérence des idées. Quand c'est possible, il sélectionne les informations essentielles et laisse de côté les idées accessoires (non essentielles) [Giquel 1990].
– **Synthèse de l'information**
  Les informations sélectionnées doivent être réduites et réunies. Cette réduction peut s'opérer à l'aide des principes d'effacement (suppression) / intégration / généralisation, et/ou par l'utilisation de procédés stylistiques visant la réduction du nombre de mots et de phrases.
  La réunion des divers éléments retenus consiste à leur donner une articulation d'ensemble, à les enchaîner les uns aux autres (synthèse) [van Dijk 1983].
– **Rédaction du texte du résumé**
  Il s'agit enfin de réécrire un texte plus court que le texte initial, en adoptant un style correct, neutre et en veillant à la clarté de la rédaction. Le résumeur humain reprend les informations du texte jugées pertinentes (synthétisées) qui fait que ces informations sont mentionnées, tout en évitant les commentaires, les illustrations, etc.

Le processus de rédaction préconise la production de mots, la constitution de phrases et leur enchaînement [Blais 2008].

1.3.2 Facteurs influençant la forme du résumé

Un résumé d'un texte source peut prendre plusieurs formes différentes. Cela est équivalent à dire qu'un texte source admet une quasi infinité de résumés qui ne sont pas toujours équivalents [Masson 1998]. Ces différentes formes de résumés dépendent essentiellement de plusieurs facteurs comme : la nature du texte source, les objectifs de l'auteur, etc.
Selon Karen Spark-Jones [Jones 1993], les facteurs affectant le résumé sont :

- **La nature de la source textuelle** : le type du texte influence la forme du résumé à produire. En effet, l'épreuve de rappel sur des sujets montre bien quelle information a été mémorisée en raison de sa pertinence, et ici la pertinence est bien liée à la nature du texte source.
- **Le rôle que le résumé doit avoir** : la fonction (information, alerte, critique, promotion, etc.) que va jouer le résumé influence son type. Aussi, l'audience (spécialiste, non-spécialiste) à laquelle est destiné le résumé affecte la forme du résumé.
- **La forme physique du résumé** : le résumé peut être fourni sous forme d'une liste de mots-clés, de sommaire, d'un texte complet, d'une séquence audio, etc. La langue dans laquelle le résumé doit être rédigé peut être la même ou différente de celle du texte source. Ainsi, l'activité résumante peut donner lieu à des produits de nature différente en fonction d'un certain nombre de paramètres. Dans les sections suivantes, nous allons détailler les types de résumés, les types d'utilisateurs du résumé et les caractéristiques inhérentes à un résumé.

1.4 Différents types du résumé textuel

Les résumés textuels peuvent être classés en fonction de leurs contenus et leurs objectifs afin qu'ils agissent correctement sur le lecteur en satisfaisant ses attentes initiales.
Le résumé peut avoir ainsi différentes visées suivant l'utilisation que l'on veut en faire et suivant le besoin préalable auquel on cherche à répondre. Plusieurs auteurs ont proposé leurs propres classifications. Ces classifications catégorisent les principaux types de résumés à partir de la fonction qu'ils jouent.
Nous exposons, dans ce qui suit, des définitions des différents types de résumés envisagés. On précise toutefois que ces types de résumés ne sont pas indépendants les uns des autres, ils possèdent des propriétés communes comme la concentration importante d'informations proportionnellement à un texte de même taille.

D'après [Minel 2002a] et Hasler [Hasler 2007], on peut distinguer les types de résumé suivants :

- *Résumé informatif* : le résumé informatif fournit un ensemble d'informations permettant de donner un large panorama du contenu d'un texte. En effet, le résumé final tend à contenir toutes les informations pertinentes du texte original. Pour cela, l'ensemble des principaux sujets doit être rapporté. Ainsi, les sujets principaux qui sont rappelés dans le résumé sont répartis de manière fidèle par rapport à l'organisation initiale afin de donner un juste aperçu du texte source.
- *Résumé indicatif* : le résumé indicatif a pour fonction de fournir au lecteur suffisamment d'informations pour qu'il puisse juger s'il doit consulter ou non le document source. Le résumé indicatif fait intervenir la notion de thématisation, c'est-à-dire, qui ne reprend que les thèmes développés dans le document source sans en prendre les commentaires. Ce type de résumé peut être apparenté à une table des matières [Aït El Mekki 2004]. Pour cela, le résumé indicatif contient seulement des éléments partiels par rapport au résumé informatif, mais surtout des éléments pertinents en vue de répondre à sa fonction. Le résumé indicatif est très souvent utilisé dans les fonds documentaires car il est bien adapté à la description des documents un peu longs : il donne un bref aperçu au lecteur de ce qu'il peut trouver dans le document. De ce fait, le résumé indicatif est utile aussi pour les articles courts car il donne un aperçu immédiat du contenu.
- *Résumé d'opinion* : le résumé d'opinion a pour but de présenter et de repérer les citations, les jugements et les opinions. Ainsi, le but du résumé d'opinion n'est pas seulement de produire une synthèse de l'information contenue dans les textes, mais en outre de dégager des tendances, d'identifier les opinions exprimées [Bossard 2010].
- *Résumé de conclusions* : le résumé de conclusions est appelé parfois "le résumé Résultats" [Grize 1990] et aussi "le résumé récapitulatif" [Charolles 1989]. Ainsi, le résumé de conclusions ou résultats est défini comme "un bref exposé dans un document (généralement placé à la fin de ce document) et qui a pour but de compléter l'orientation du lecteur qui a étudié le texte précédent". Pour cela, ce type de résumé reprend uniquement les résultats et les conclusions présentés dans le texte source.
- *Résumé critique* : ce résumé ne contient pas uniquement les informations pertinentes mais il contient aussi les opinions du résumeur (autre personne que l'auteur du texte). Le résumeur évalue d'une façon critique la qualité et les assertions majeures exprimées dans le document originel. Pour cela, ce type de résumé combine le condensé du texte source avec les apports critiques sur le contenu de ce texte.
- *Résumé synthétique* : il consiste à emprunter certains termes du texte source et à en faire une interprétation donnant lieu à un nouveau texte qui n'est pas un sous-ensemble du texte-source.
- *Résumé scolaire* : le résumé scolaire obéit à des critères de construction précis comme

un taux de réduction normé, l'interdiction d'emprunts ou le recours systématique aux synonymes. En effet, ce type de résumé a été conçu dans un objectif pédagogique afin d'évaluer et de vérifier les capacités d'un élève dans l'analyse et la compréhension d'un texte et dans la rédaction. Le résumé scolaire doit être ainsi fidèle au texte original, en conservant les grandes lignes et la structure générale du sujet traité.

1.5 Types d'utilisateurs des résumés

Il n'existe pas de résumé sans fonction particulière. Il est donc clair que la production de résumés ne peut pas être envisagée sans avoir à l'esprit les éléments qui gouvernent les relations qui existent entre le résumé et le lecteur [Masson 1998].
Cependant, il est assez difficile de trouver des critères pour définir a priori les utilisateurs. On peut les classer de plusieurs façons différentes, par exemple :

- Les spécialistes du (ou des) domaine(s) abordé(s) dans le texte source,
- Les non-spécialistes de ces domaines.

Une dichotomie de ce type pose directement le problème du niveau de compréhension du lecteur et du niveau de difficultés du résumé. Il est assez clair qu'un spécialiste aura besoin des termes techniques précis liés au domaine, alors qu'un non-spécialiste sera plutôt rebuté par ces mêmes formes. Le niveau de connaissance, de langage, et l'abstraction d'un résumé constitué par extraction de phrases étant le même que celui du texte source. Masson pense qu'il vaut mieux réserver ce type de résumé à l'usage des spécialistes (ou du moins personnes bien informées). Il présente aussi dans le cas d'un texte source "simple à comprendre" de type vulgarisation pour tout public, cette remarque devient caduque.
Les résumés de type synthèse seront, quant à eux, compréhensibles par tout le monde ; ils peuvent donc être destinés aux non spécialistes comme aux spécialistes. Cependant, ce type de résumé pourra dans certains cas ne pas plaire aux spécialistes qui attendent du résumé un contenu informationnel dense.

1.6 Caractéristiques des résumés

Un résumé doit être doté d'un certain nombre de caractéristiques : la concision, la couverture, la fidélité, la cohésion et la cohérence. Nous exposons ces caractéristiques telles qu'elles ont été définies dans [Masson 1998].

1.6.1 Concision

La **concision** est à relier directement au taux de réduction qui est le rapport entre la longueur du texte source et celle du résumé. D'une manière générale, le taux de réduction est proportionnel au caractère restrictif du (ou des) critère(s) employé(s) pour générer le résumé. Par exemple, un critère très restrictif du genre "ne retenir que les expressions conclusives et de résultats" produira des résumés plutôt courts. À l'inverse, un critère plus vague et beaucoup moins restrictif du genre "faire un résumé informatif" conduira dans la plupart des cas à des résumés plus longs. Il est aussi important de retenir qu'un processus de résumé faisant intervenir la reformulation permet de générer des résumés considérablement plus courts que ceux produits sans processus de reformulation [Grize 1990].

1.6.2 Couverture

La **couverture** est en quelque sorte le rapport entre le nombre de thèmes ou d'éléments présents dans le texte source et ceux présents dans le résumé. La nature des éléments est fonction du type de résumé considéré : pour un résumé indicatif, on retiendra uniquement les thèmes abordés ; dans un résumé résultat, on s'intéressera principalement aux expressions conclusives et aux résultats. Dans le cas d'un résumé informatif, c'est-à-dire, une réduction "à l'identique" du texte, la couverture est plus difficile à déterminer.

1.6.3 Fidélité

La **fidélité** est aussi un critère important pour caractériser un résumé. Elle représente la relation de similarité objective existant entre le résumé et le texte source. C'est en quelque sorte une mesure de la qualité globale du résumé. La notion de fidélité intègre, comme composante, la couverture. En règle générale, un résumé ayant une couverture correcte sera assez fidèle au texte source.

1.6.4 Cohésion et cohérence

Les deux derniers critères que nous exposerons comme définissant un résumé sont intimement liés à la notion de texte elle-même. Il s'agit de la **cohésion** et de **cohérence**. La cohésion peut être vue comme le résultat de l'application de mécanismes visant à maintenir une unité référentielle (mécanisme d'anaphore) et argumentative (emploi des connecteurs). La cohérence quant à elle découle plus de la bonne application des mécanismes rhétoriques et thématiques (suivi du thème), qui rendent un texte intelligible.

1.7 Motivations pour l'automatisation du résumé

En évoquant l'automatisation de l'activité résumante, signalons que cette tâche est potentiellement très pratique, car elle fournit une solution partielle pour le problème d'actualité, à savoir l'explosion de l'information. Ainsi, l'idée d'affecter la tâche de résumer des documents à une machine plutôt qu'à des humains est alors née en raison des économies que pourrait engendrer cette automatisation. La quantité croissante de documents (textes techniques et scientifiques, articles de presse, etc.) associée à la composante temporelle qui devient encore plus restreinte au point où elle ne suffit plus pour traiter cette abondante masse d'information, voire même pour y accéder tout simplement. L'apparition d'Internet et l'explosion de documents sous formats numériques (en ligne et hors ligne), la montée en puissance de la recherche d'informations depuis les années 1990 a également entraîné dans son sillage le résumé automatique qui en est devenu une branche particulière. Malgré une certaine stagnation du domaine depuis certaines années, dans ses avancées théoriques comme dans ses applications, l'enjeu de produire des résumés automatiquement reste toujours une activité incontournable afin d'aider le chercheur à décider l'ensemble des documents qu'il retiendra.

1.8 Conclusion

Dans ce chapitre, nous avons étudié la question à aborder dans cette thèse à savoir le résumé. Nous nous sommes intéressés au processus humain visant la production d'un résumé, aux types et aux caractéristiques de ce dernier. Nous avons aussi focalisé les types d'utilisateurs des résumés. En effet, en visant une automatisation de l'activité résumante, il faut nettement que le système prenne en compte les intérêts de l'utilisateur plutôt que celles du concepteur.
Dans le chapitre suivant, nous allons présenter un aperçu sur les principales approches et méthodes de résumé automatique.

Chapitre 2

État de l'art sur les approches de résumés automatiques

Sommaire

2.1 Introduction 18
2.2 Approche numérique 18
2.2.1 Méthodes axées sur les calculs statistiques 19
2.2.1.1 Fréquence des mots 19
2.2.1.2 Mots du titre 20
2.2.1.3 Position des phrases 20
2.2.1.4 Statistiques de co-occurrences lexicales 21
2.2.1.5 Expressions indicatives 21
2.2.2 Méthodes fondées sur l'apprentissage 22
2.2.2.1 Définition de l'apprentissage automatique 22
2.2.2.2 Modèle de résumé automatique basé sur l'apprentissage supervisé 23
2.2.2.3 Modèle de résumé automatique basé sur l'apprentissage non-supervisé / semi-supervisé 27
2.2.2.4 Synthèse sur les méthodes de résumé par apprentissage 28
2.2.3 Discussion des méthodes numériques 29
2.3 Approche symbolique 29
2.3.1 La grammaire de Montague 30
2.3.2 La théorie de la structure rhétorique 31
2.3.3 La théorie de la représentation discursive 34
2.3.4 Techniques de la théorie de la représentation discursive segmentée (SDRT) 36
2.3.5 Techniques de l'exploration contextuelle 40
2.3.6 Discussion des méthodes symboliques 41
2.4 Approche hybride 43
2.5 Conclusion 43

2.1 Introduction

L'idée de produire des résumés de manière automatique n'est pas nouvelle, la première initiative d'aborder ce sujet a été proposée par Lunh en 1958 [Luhn 1958]. Jusqu'aux années 80, les investigations sur ce domaine se sont effectuées d'une façon ininterrompue. Ce n'est qu'à partir des années 90, années de l'essor de l'Internet et de la croissance exponentielle de la masse d'information disponibles en format électronique, que la problématique d'offrir des outils informatiques de visualisation rapide des textes, par exemple des résumés automatiques, afin que l'utilisateur puisse évaluer la pertinence d'un document vis-à-vis de l'information cherchée, a pris toute sa valeur. En effet, depuis ces années, ce sujet de recherche est couramment mentionné dans les thèmes retenus pour les grandes conférences en Traitement Automatique des Langues Naturelles (TALN) ou en Recherche d'Information (RI). Il a même fait l'objet d'un intérêt particulier exprimé à travers de nombreux colloques internationaux comme EACL (European Chapter of the Association for Computational Linguistics), NAACL (North American Chapter of the Association for Computational Linguistics), TALN (Traitement Automatique des Langues Naturelles), et de conférences d'évaluation comme TREC (Text REtrieval Conferences), DUC (Document Understanding Conferences) spécifiques aux résumés automatiques de textes ou TAC (Text Analysis Conference).
Ainsi, différentes méthodes ont été développées au cours des trente dernières années pour produire automatiquement un résumé à partir d'un texte d'origine [Minel 2002a]. Ces méthodes peuvent être classées en deux groupes : l'approche *numérique* fondée sur les techniques à base des scores/poids, et l'approche *symbolique* fondée sur les techniques purement linguistiques basées sur une étude sémantique.
Ce chapitre a pour objet de rendre compte d'une recherche bibliographique concernant notre domaine d'étude, à savoir le résumé automatique de textes. Dans un premier temps, nous présentons l'approche numérique développée dans le cadre de l'extraction automatique, et qui est basée sur des méthodes statistiques et des méthodes d'intelligence artificielle à savoir l'apprentissage. Nous présentons, dans un deuxième temps, l'approche symbolique de génération de résumé, qui a influencé notre travail. Nous achèverons ce chapitre par bref aperçu sur l'approche hybride et par une synthèse des diverses méthodes de résumé automatique.

2.2 Approche numérique

L'approche numérique repose sur des méthodes dont le but est de trouver un sous-ensemble du texte source qui est indicatif de l'essence de son contenu. Les méthodes issues de ce paradigme sont des méthodes empiriques qui ont pour objectif l'identification des unités qui influencent l'importance de la phrase.

L'approche numérique se confine typiquement à une affectation des scores/poids aux mots et par la suite aux phrases en se basant sur des règles spécifiques. Ces règles sont principalement relatives à l'identification d'indicateurs ou critères pour l'importance de chacune des phrases du texte source.
Les critères d'attribution de l'importance, aux mots du texte source, peuvent être statistiques en considérant la fréquence des occurrences ou non statistiques en considérant uniquement l'occurrence de signaux linguistiques tels que par exemple les mots du titre. Un poids pour chaque phrase est donc obtenue moyennant une fonction des scores de ses constituants en mots. La génération du texte du résumé est donc réduite à la concaténation des phrases qui présentent les scores les plus élevés dans l'ordre de leur occurrence dans le texte originel. Le principe général des techniques d'extraction contourne donc les difficultés classiques du traitement automatique de la langue naturelle, il révèle une efficacité opérationnelle et assure l'indépendance du système vis-à-vis du domaine et du corpus utilisé. Les principales méthodes utilisées pour identifier les indicateurs de l'importance des phrases et qui ont été utilisées par plusieurs chercheurs, sont exposées dans ce qui suit.

2.2.1 Méthodes axées sur les calculs statistiques

2.2.1.1 Fréquence des mots

Cette méthode est considérée parmi les premières méthodes expérimentées dans le domaine de résumé automatique. Elle a été développée par Luhn en 1958 [Luhn 1958]. Elle se base sur le fait que l'auteur utilise, pour exprimer ses idées clés, quelques mots clés qui ont tendance à être récurrents dans le texte. En effet, cette suggestion repose sur l'hypothèse qu'un auteur met normalement l'accent sur un aspect d'un sujet en répétant certains mots qui lui sont relatifs. Les mots à haute fréquence sont donc indicatifs du contenu du document et sont considérés comme positivement représentatifs [Ellouze 2004].
"La procédure suggérée revendique sur le principe que les mots de haute fréquence dans un document sont les mots importants" [Luhn 1958]. Les phrases importantes sont celles qui renferment les mots assez fréquemment employés dans le texte.
Luhn cumule au critère de fréquence un deuxième critère qui est la proximité[1]. Deux mots appartiennent à un même groupe de mots clés ou "*pleins*" si la distance entre les deux mots ne dépasse pas quatre ou cinq mots non significatifs ou "*vides*". Dans cet ordre d'idées, le poids attribué à chaque phrase dépend de la richesse de la phrase en mots appartenant à la liste des mots clés (un mot clé est un mot qui possède une fréquence d'apparition qui dépasse un seuil préétabli). En effet, Luhn a basé le score complet de la phrase sur les groupes des mots-clés contenus dans chaque phrase. La représentativité de la phrase peut être évaluée en calculant la

1. Proximité : (du latin proximus, proche) nom féminin, Voisinage immédiat. - À proximité de : près de.

moyenne des valeurs de représentativité des éléments non isolés de la phrase.
L'étape finale consiste à sélectionner les n premières phrases qui ont les poids les plus élevés et qui sont considérées comme les plus pertinentes. Le procédé d'extraction des n premières phrases destinées à construire le résumé est défini soit à priori une valeur (seuil) à laquelle la valeur de représentativité (scores/poids) doit être supérieure, soit on fixe une longueur maximale de l'extrait en nombre de phrases et on ne retient que celles ayant les valeurs représentatives les plus importantes [Ellouze 2004].

2.2.1.2 Mots du titre

Étant donné que le titre est l'expression la plus significative et qui résume le mieux un document en quelques mots, on peut dire que la phrase qui ressemble le plus au titre est la plus marquante du document, du fait que les principaux thèmes sont véhiculés en général dans les titres et les sous-titres [Douzidia 2004a].
Dans ce cas, on considère les mots du titre du texte comme des mots clés d'indexation. La liste d'index de termes est créée pour le document avant le processus de pondération des phrases. Les termes candidats sont sélectionnés à partir du titre, des sous-titres du document. Edmundson [Edmundson 1969] propose d'assigner des poids lourds pour les mots pleins du titre du document ainsi que pour les mots pleins des titres et sous-titres des sections. Par ailleurs, les travaux de Ben Hamadou [Ben Hamadou 1995] vont dans la même direction. Les auteurs proposent d'assigner des poids élevés pour les co-occurrences de mots de titres et des poids faibles pour les mots isolés. Par conséquent, les phrases contenant des co-occurrences de mots de titres auront des poids plus lourds que celles contenant des mots isolés [Ellouze 2004].

2.2.1.3 Position des phrases

Cette méthode a été introduite par Edmunson en 1969 [Edmundson 1969] pour compléter la méthode de distribution de termes qu'il a appelée "*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 [Ishikawa 2001].
Cette méthode considère que la position de la phrase dans le texte détermine le degré de son importance. Dans cet ordre d'idées, cette méthode suppose que la position d'une phrase dans un texte indique son importance dans le contexte. Les premières et les dernières phrases d'un paragraphe par exemple, peuvent transmettre l'idée principale et devraient donc faire partie du résumé.
Comme variante de cette méthode, on peut citer la méthode *Lead* [Ishikawa 2001] ; c'est une méthode qui détermine les phrases importantes en extrayant celles qui sont en tête.
Cette méthode est efficace pour résumer les articles de journaux, puisque les phrases importantes

ont tendance à apparaître dans les premières phrases de l'article.
L'inconvénient de cette méthode est qu'elle dépend de la nature du texte à résumer ainsi que du style de l'auteur [Douzidia 2004a].

2.2.1.4 Statistiques de co-occurrences lexicales

Cette méthode se situe dans la même lignée de celle proposée par [Hahn 1998a] sauf que le calcul de la fréquence des mots clés ou "*pleins*" doit tenir compte du corpus auquel appartient le texte à résumer. En effet, il faut noter que lorsqu'on traite un texte qui appartient à un corpus sur un sujet particulier (informatique, économie, finance, etc.), d'autres mesures doivent être prises en considération. En fait, 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 "systèmes distribués" soient seulement spécifiques à un sous-ensemble de documents. Dans ce cas, la fréquence de chaque mot doit être normalisée [Ellouze 2004].
La fréquence normalisée du mot i est donnée par la formule suivante [Salton 1989], [Salton 1997]

$$fr_i = tf_i * idf_i \tag{2.1}$$

- fr_i : fréquence normalisée du mot i.
- tf_i : fréquence du mot i dans le document à résumer.
- idf_i : fréquence inverse du mot i.

avec :

$$idf_i = log\left(\frac{N}{dtf_i}\right) \tag{2.2}$$

Où N est le nombre de mots dans la collection de référence et dtf_i est le nombre de phrase contenant le mot i.
Après avoir calculé la fréquence de chaque mot, un poids est attribué à chaque phrase. Le résumé généré est alors produit en affichant les phrases les plus "pesantes" [2] du document source.

2.2.1.5 Expressions indicatives

Dans cette méthode, la pondération des phrases dépend de deux types d'expressions : les expressions "bonus" et les expressions "*stigma*" [Edmundson 1969].
Les expressions bonus contiennent principalement des superlatifs et des mots pleins tels que

2. Phrases ayant les sores les plus grands.

"cet article présente", "dans ce rapport nous proposons", "en conclusion", elles indiquent que l'auteur est en train d'annoncer le thème général de son document et en conséquence elles augmentent le score de la phrase qui les contient. En contre partie, les expressions "*stigma*" contiennent principalement des anaphores et des mots ayant une valeur diminutive tels que "par exemple", "impossible", "infaisable", elles pénalisent donc le poids de la phrase.
Le poids final de chaque phrase est calculé en sommant les poids des indicateurs trouvés parmi les mots qui la constituent [Saggion 2000].

2.2.2 Méthodes fondées sur l'apprentissage

Pour commencer cette étude sur les méthodes de résumé automatique basées sur l'apprentissage, nous proposons d'abord quelques définitions d'apprentissage automatique. Ces définitions vont nous permettre d'identifier la fonction, l'objectif, et le résultat attendus de ces méthodes.

2.2.2.1 Définition de l'apprentissage automatique

Selon Amini [Amini 2001] et Young-Min [Young-Min Kim 2010] "l'apprentissage est une façon d'exploiter les caractéristiques de corpus et l'interaction utilisateur. La prise en compte de ces deux aspects permet d'améliorer significativement la quantité des phrases extraites qui forme l'extract".
D'après Simon [Simon 1983] "L'apprentissage dans un système est indiqué par les changements qu'il subit. Ces changements sont adaptatifs dans le sens où ils rendent possible au système de réaliser une même tâche, ou des tâches tirées d'une même population, d'une façon plus efficace et plus efficiente la prochaine fois qu'elle sera réalisée".
Pour Osório [Osório 1998] "L'apprentissage automatique se fait par des outils qui permettent d'acquérir, élargir et améliorer les connaissances disponibles au système". En général, l'apprentissage implique des processus d'adaptation et de modification des structures de contrôle et/ou de représentation de connaissances du système en question.
Nikolopoulos [Nikolopoulos 1997] présente l'apprentissage automatique comme une alternative prometteuse pour améliorer le processus d'acquisition de connaissances. Les systèmes experts de première génération, du fait des problèmes liés à l'acquisition de connaissances, ont évolué vers des systèmes dits de deuxième génération qui ont intégré, parmi d'autres techniques, des outils d'apprentissage automatique.
Dans l'apprentissage automatique on cherche à acquérir des règles générales qui représentent les connaissances obtenues à partir d'exemples. En effet, le principe de l'apprentissage automatique consiste à générer dans un premier temps des classifieurs qui servent ensuite à classer les objets (textes, phrases, formes graphiques, etc.). Contrairement aux autres méthodes sans apprentissage, les méthodes d'apprentissage nécessitent systématiquement un corpus d'entraî-

nement permettant une classification ultérieure sur les données à classer.
Généralement, la plupart des méthodes reposent sur l'apprentissage supervisé. Ce dernier nécessite un corpus annoté manuellement. Or l'annotation manuelle au niveau entités textuelles est très coûteuse en temps et de plus, trier les phrases est intrinsèquement difficile du fait qu'il n'existe pas de caractérisation de ce qu'est un bon résumé [Minel 2002b].
Dans la suite de cette section, nous présentons les trois types d'apprentissage automatique à savoir : l'apprentissage supervisé, l'apprentissage semi-supervisé et l'apprentissage non-supervisé (ou par renforcement).

2.2.2.2 Modèle de résumé automatique basé sur l'apprentissage supervisé

En apprentissage *supervisé*, l'objectif principal consiste à optimiser un critère qui mesure la similarité entre les sorties que produit le système et les sorties désirées [Amini 2003].
Récemment, différents auteurs ont commencé à s'intéresser aux techniques d'apprentissage pour effectuer des résumés automatiques de textes. Ces techniques permettent de s'adapter au corpus traité ou aux demandes particulières de l'utilisateur. Toutefois, toutes les méthodes proposées pour ce domaine reposent sur l'apprentissage supervisé. Afin, d'apprendre ces méthodes, il est nécessaire d'avoir un corpus étiqueté, généralement manuellement, de toutes les entités textuelles des documents d'entraînements. L'étiquetage manuel au niveau entités textuelles est très coûteux en temps et infaisable pour la plupart des applications réelles [Minel 2002b].
Les systèmes construits ces dernières années sont de type supervisé. Les systèmes basés sur l'apprentissage supervisé nécessitent un corpus d'entraînement ou d'apprentissage composé de textes sources et de leurs résumés. Or cette exigence peut ne pas être remplie et ce pour plusieurs raisons [Amini 2003] :

- Le coût croissant d'élaboration des résumés de qualité.
- L'intérêt de certains textes peut être très limité dans le temps, ce qui rend le coût de production d'un résumé prohibitif.
- Les normes de production des textes peuvent évoluer dans le temps, ce qui entraîne une modification de l'importance de certains critères structurels appris par un système supervisé.
- Des nouvelles formes sémiotiques, notamment sur la toile, apparaissent de plus en plus, rendant ainsi aléatoire toute tentative de figer la pondération entre les critères d'apprentissage.

A partir d'un corpus d'apprentissage ou d'entraînement composé d'un ensemble de textes sources et de leurs résumés correspondants, ces résumés vont être considérés dans la phase d'apprentissage comme des références.
Les systèmes supervisés, basés sur l'extraction, se composent généralement de deux étapes

illustrées dans la figure 2.1.

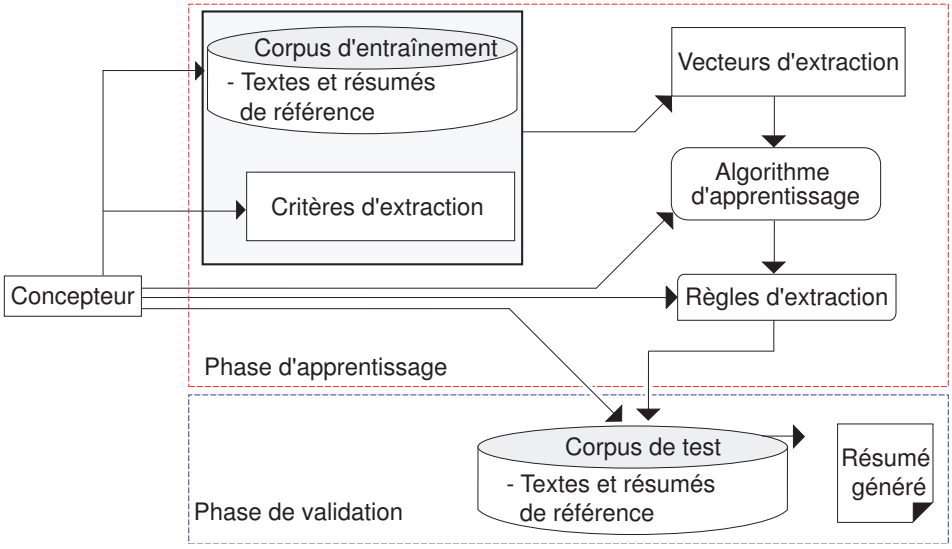

FIGURE 2.1 – Principe des systèmes à apprentissage supervisé [Minel 2002b].

Comme le montre la figure 2.1, les systèmes à apprentissage supervisé utilisent un corpus d'entraînement composé d'un ensemble de textes et de leurs résumés de référence. En effet, la première étape, la *phase d'apprentissage*, consiste à choisir les critères qui déterminent l'extraction d'une phrase. Ces critères combinent des informations positionnelles (e.g. la phrase est située dans le premier paragraphe), lexicales (e.g. la phrase contient des mots fréquents dans le texte), structurelles (e.g. la phrase contient des mots présents dans le titre du texte), etc.

Ceci implique d'analyser chaque phrase, pour tous les textes du corpus, et de les annoter au regard des critères d'extraction qui sont considérés comme pertinents. Toutefois, il faut noter que ces critères d'extraction doivent impérativement pouvoir être calculés afin d'éviter une annotation manuelle dans la deuxième étape (*phase de validation*).

Le tableau 2.1 présente des exemples de critères utilisés par Mani et Bloedron [Mani 1998]. En fait, l'objectif de la phase de validation consiste à choisir des bons critères d'extraction. Ainsi, la fin de cette deuxième étape, aboutit à la construction d'un ensemble de vecteurs d'extraction $V(d_i, p_j)$ relative à chaque phrase p_j d'un document d_i pris de l'ensemble $D = (d_i, \ldots, d_m)$ de documents corpus d'entraînement. Ensuite, les phrases de chaque document d_i de la base d'apprentissage, sont regroupées en deux classes : celles qui doivent appartenir au résumé extractif (classe +1), et les autres (classe -1). Ainsi, pour chaque document d_i, on possède un ensemble P_j de couples (vecteur, étiquette de classe), chaque vecteur $V(d_i, p_j)$ étant représentatif d'une phrase p_j du document d_i.

Pour la première étape d'apprentissage elle utilise un algorithme d'apprentissage pour produire

TABLEAU 2.1 – Exemple de critères d'extraction utilisés pour l'apprentissage [Mani 1998]

	Critères	Espaces de valeurs	Commentaires
Critères positionnels	Sent-loc-para	1,2,3	La phrase est placée au début, au milieu, ou dans le dernier tiers du paragraphe.
	Para-loc-section	1,2,3	La phrase est placée au début, au milieu, ou dans le dernier tiers de la section.
	Sent-spécial-section	1,2,3	Égale à 1 si la phrase est placée dans l'introduction, à 2 si la phrase est placée dans la conclusion, à 3 dans les autres cas.
	Depth-sent-section	1,2,3,4	Égale au rang de la section dans laquelle la phrase est placée.
Critères lexicaux	Sent-in- highest-tf	Booléen	Le score tf de la phrase est le plus élevé.
	Sent-in-height-tf.idf	Booléen	Le score tf*idf de la phrase est le plus élevé.
	Sent-in-highest-title	Booléen	Le nombre de titres dans lesquels apparaissent des termes de la phrase est le plus élevé.
	Sent-in-hightest-pname	Booléen	Le nombre d'entités nommées qui apparaissent des termes de la phrase est le plus élevé.
Critères de cohésion	Sent-in-highest-syn	Booléen	Le nombre de liens de synonymie est le plus élevé.
	Sent-in-highest-co-occ	Booléen	Le nombre de liens de cooccurrence est le plus élevé.

des règles d'extraction. Les vecteurs V constituant l'entrée à partir de laquelle les règles sont construites [Minel 2002a]. Une base d'apprentissage est alors crée :

$$P = \bigcup_{i=1}^{m} P_i = \left\{ \left(V\left(d_i, p_i\right), y_i^j \right) | i \in \{1, ..., m\}, j \in \{1, ..., |P_i|\} \right\} \tag{2.3}$$

Où chaque

$$y_i^j \in \{-1, +1\} \tag{2.4}$$

Après entraînement, l'extraction des phrases pour le résumé d'un nouveau document est effectuée en ordonnant les phrases selon le score renvoyé par le classifieur, qui est une estimation de la probabilité de la phrase d'appartenir au résumé [Usunier 2006].
Pour cela, plusieurs chercheurs comme Kupiec [Kupiec 1995], Amini [Amini 2003] et Lin [Lin 1998] ont utilisé des algorithmes d'apprentissage basés sur une fonction de score f

pour minimiser l'erreur de classification :

$$R_{|P|}(F,P) = \frac{1}{|P|}\sum_{i=1}^{m}\sum_{j=1}^{|P_i|} L\left(f\left(d_i,p_j\right),y_j^i\right) \tag{2.5}$$

Dans la deuxième étape (i.e.*la phase de validation*) chaque phrase du document d'origine est analysée en fonction des critères choisis par la première étape, puis comparée au résumé de référence.

Il faut signaler dans cette étape que la comparaison est effectuée en utilisant un *calcul de similarité* dans le cas où le résumé de référence n'est pas composé de phrases du texte d'origine. Toutefois, le calcul de similarité peut être réalisé par plusieurs façons. En général, le calcul de similarité peut être modifié selon les critères qui déterminent l'extraction d'une phrase. Ces critères d'extraction sont définis par la première étape.

Selon Usunier et al. [Usunier 2006] les critères qui peuvent modifier le calcul de similarité sont soit des critères d'informations positionnelles, lexicales ou structurelles.

$$Sim_1(q,p) = \frac{\sum_{w\in q\bigcap p} tf(w,q) * tf(w,p) * idf(w)^2}{\|p\| * \|q\|} \tag{2.6}$$

Avec :

- $tf(w,p)$ est le nombre d'occurrences du mot w dans la phrase p,
- $tf(w,q)$ est le nombre de fois que le mot apparaît dans la liste du mots qui forme le critère d'extraction
- $idf(w) = \ln\left(\frac{|D|}{df(w)} + 1\right)$ est l'inverse document frequency du mot w, où
  - $|D|$ représente le nombre de documents dans la collection
  - $df(w)$ est le nombre de documents de la collection dans laquelle le mot w apparaît au moins une fois.
- et $\|P\| = \sqrt{\sum_{w\in p}(tf(w,x) * idf(w))^2}$

Cette formule 2.6 est issue de la mesure de similarité $tf * idf$ en recherche d'information. Elle représente le cosinus de l'angle entre la phrase p et le critère q lorsqu'ils sont tous les deux représentés dans l'espace des mots, avec la valeur $tf(w,x) * idf(w)$ pour la dimension correspondant au mot w, où $x = p$ ou $x = q$.

Dans la dernière étape de validation, le système basé sur l'apprentissage supervisé utilise le jugement d'un utilisateur pour corriger ses erreurs de décision. En effet, cette étape a comme but de valider les règles produites en les appliquant sur un corpus de test, et en comparant les résumés produits automatiquement avec les résumés de référence.

2.2.2.3 Modèle de résumé automatique basé sur l'apprentissage non-supervisé / semi-supervisé

Les systèmes de résumé basés sur l'apprentissage *non-supervisé* ou *semi-supervisé* se proposent de fournir une solution pour construire des systèmes par apprentissage sans disposer de corpus d'entraînement ou en s'appuyant sur un corpus d'apprentissage de petite taille. Dans cette approche on va plus loin dans le sens de l'automatisation car le système n'a pas besoin de l'interaction utilisateur pour apprendre (voir figure 2.2). Il faut noter, qu'il n'est pas possible, pour le concepteur du système comme c'est le cas pour les systèmes supervisés, de modifier la fonction de vraisemblance. En fait, ces systèmes sont plutôt du type "*Boîte noire*". Toutefois, il existe deux approches pour la classification en monde *semi-supervisée* [Amini 2003].

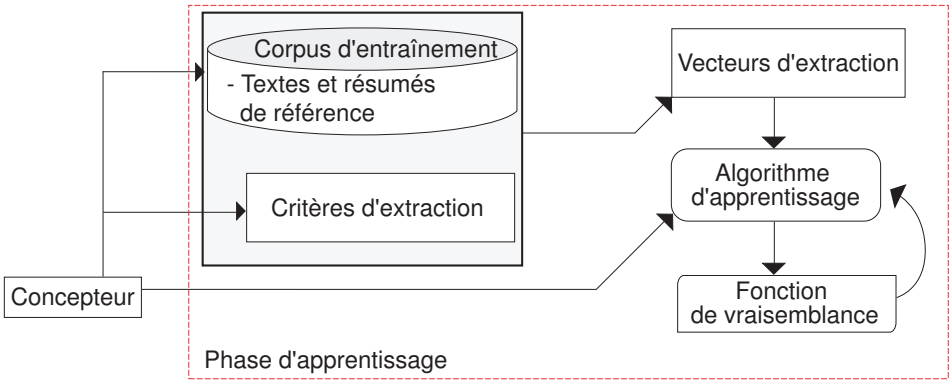

FIGURE 2.2 – Principe des systèmes à apprentissage semi-supervisé [Amini 2001]

La première approche de classification est dite *générative* et la seconde est dite *discriminante*. Dans la *classification générative*, on tente de modéliser la distribution des données en maximisant la vraisemblance du mélange, alors que dans la *classification discriminante* on tente plutôt de calculer directement la pertinence des phrases par rapport au résumé, sans modéliser les données.

La plupart des travaux réalisés dans ce dernier cas, proposent d'appliquer une estimation de maximum de vraisemblance et d'adapter l'algorithme connu sous le nom d'algorithme *Espérance-Maximisation* (EM) pour prendre en compte des données étiquetées et non étiquetées. Il est à noter que le critère de vraisemblance est classifiant, ce qui a pour avantage qu'aucune hypothèse n'est faite sur les données, ce qui est en général justifié [Dempster 1977].

En effet, cette maximisation est effectuée en faisant appel à une classe générale de procédures itératives de l'algorithme EM. Ainsi, l'algorithme EM alterne des étapes d'évaluation de l'espérance (E), où l'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées, et une étape de maximisation (M), où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E. On utilise ensuite

les paramètres trouvés en M comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi, les vecteurs indicateurs de classes. Nous allons décrire comment cet algorithme a été appliqué au domaine du résumé automatique, en se basant sur les travaux de Amini [Amini 2001].
Le système dispose au départ d'un ensemble de taille restreinte D_1composé des phrases d'une dizaine de texte, pour lesquels les résumés sont connus et d'un nombre de phrases issues d'un plus grand nombre de textes pour lesquels aucun résumé n'est disponible. On peut considérer que le calcul de la fonction de vraisemblance consiste à étiqueter les phrases des textes avec deux étiquettes conserver ou *éliminer*. L'algorithme utilisé est décrit de manière informelle ci-dessous.

Algorithme 1 Algorithme d'apprentissage semi-supervisé [Amini 2001]

1: **Entrée** : Un ensemble de phrases étiquetées D_1 et des phrases non étiquetées D_u.
2: Apprentissage d'un classifieur en mode supervisé sur D_1.
Sur l'ensemble des données composées de D_1 union D_u **Faire** :
3: **while** (les données étiquetées de D_u varient) **do**
On déduit un nouvel étiquetage des phrases Réapprendre le classifieur en mode supervisé sur D_1 union D_u
4: **Sortie** : Ensemble des phrases étiquetées

Comme illustré par cet algorithme, on voit que le système ré-applique à chaque itération son calcul de la fonction de vraisemblance à partir de ses propres sorties, calculées lors de l'itération précédente. Comme il existe plusieurs chemins possibles pour calculer cette fonction, chaque itération produit une fonction de vraisemblance différente qui tend à converger vers un maximum local. Ce principe est utilisé depuis de nombreuses années dans les systèmes de réseaux de neurones pour réaliser des classifications dans des ensembles de données.
Les concepteurs de cette approche [Amini 2003] ont d'ailleurs utilisé un système basé sur l'algorithme d'apprentissage de neurones pour résumer un corpus d'un million de dépêches de presse. Les critères d'extraction utilisés sont les critères lexicaux (la phrase contient des mots fréquents dans le texte) classiquement utilisés comme le paramètre $tf * idf$.

2.2.2.4 Synthèse sur les méthodes de résumé par apprentissage

Les techniques d'apprentissage ne sont pas une particularité des systèmes de résumé automatique. De ce fait, toutes ces techniques ont fait l'objet, dans les années 1990, de nombreuses recherches, et ont suscité de nombreux espoirs.
Les résultats n'ayant pas été à la hauteur des espérances, ces techniques ont peu à peu perdu de leur attrait, tout au moins dans la communauté européenne. Depuis quelques temps, et particulièrement dans le domaine des systèmes de résumé automatique, cette approche fait l'objet

d'un net regain d'intérêt [Stéphanie Weiser 2008].
Les résultats du système supervisé restent supérieurs à ceux des systèmes semi-supervisés ou non supervisés. Mais les différences ne sont pas très importantes au regard du coût de développement de chaque système [Amini 2010].
Notons aussi que la performance des systèmes basés sur l'apprentissage dépend étroitement des choix effectués (au niveau de la construction des corpus, des critères d'extraction, etc.) par le concepteur du système. Ils ne peuvent donc pas faire émerger des règles qui contiendraient de nouveaux critères ; dans ce sens, le terme d'apprentissage peut prêter à confusion, puisque dans le sens commun, il sous-entend la capacité, pour celui qui apprend, de produire une certaine nouveauté.

2.2.3 Discussion des méthodes numériques

Nous avons essayé de monter ci-dessus, les principales méthodes de l'approche numérique qui s'appuient sur une analyse des fréquences et de la co-occurrence des mots d'un texte. Ces méthodes numériques présentent l'avantage d'être robustes (n'importe quel texte aura un résumé). En fait, ces systèmes ont deux principaux arguments en leur faveur qui peuvent être résumés en deux points.
Un premier argument, qui est en faveur de ces systèmes, utilise l'approche numérique pourrait être le gain en temps qu'ils apportent dans le processus d'acquisition des connaissances en l'occurrence la pondération entre les critères d'extraction considérés. Cet argument est un de ceux mis en avant par les concepteurs utilisant l'approche symbolique basée sur les patrons par exemple. Le deuxième argument concerne leur parfaite adaptation à un type de textes donnés. On peut ainsi envisager qu'une entreprise qui doit analyser des articles issus d'une même source, comme la presse spécialisée ou des rapports internes rédigés en suivant des consignes rédactionnelles, puisse tirer profit de ce type de système de résumé basé sur la techniques fondées sur une approche numérique.
De notre point de vue, le principal défaut, de cette approche numérique tient à son hypothèse, implicite, de considérer la phrase comme l'élément automatique de classification. De ce fait, aucune méthode d'extraction ne prend en compte la dépendance entre phrases, dépendance qui apparaît à travers la présence d'anaphores, de connecteurs, de marques de temps, etc.

2.3 Approche symbolique

Si les études numériques ont les avantages d'être portables, applicables à tout type de corpus, et fréquemment capables d'extraire dans de bonnes proportions les résultats escomptés, ceux-ci sont toutefois produits sans explications : deux mots sont par exemple désignés comme

synonymes sans qu'aucune information additionnelle n'explique ce qui a conduit à cette décision.
L'approche symbolique de l'extraction s'oppose à l'approche numérique en particulier sur ce point.
Depuis environ deux décennies, l'utilisation d'une approche symbolique basée sur une analyse du discours et de sa structure, est imposée. En effet, l'approche purement symbolique se fait généralement par une représentation formelle des connaissances contenues dans les documents ou bien sur des techniques de reformulation [Marcus 1980], [Sitbon 2007].
De nombreux travaux sur l'analyse de discours se sont imposés. Ils partent de l'idée que la structure et la cohérence d'un texte pouvent être modélisées au moyen de relations rhétoriques ([Hobbs 1978], [Grosz 1986], [Mann 1988], [Asher 1993] [Asher 2000], [Taboada 2006], [Mathkour 2008]). Ces relations constituent un outillage précieux qui permet d'envisager l'objet linguistique, qui est le discours, sous différents angles : (i) hiérarchisent le texte en connectant et regroupant les phrases ; (ii) produisent un complément non compositionnel à la sémantique de la phrase en révélant les marqueurs qui sous-tendent la cohérence du texte ; (iii) établissent un modèle pragmatique pour représenter les organisations argumentatives et/ou intentionnelles qui charpentent un discours [Amsili 2002].
L'objet de cette section est de mettre l'accent sur quelques techniques de l'approche symbolique à savoir :

- La grammaire de Montague [Montague 1973],
- La théorie de la structure rhétorique – RST (Rhetorical Structure Theory) [Mann 1988],
- La théorie des représentations discursives – DRT (Discourse Representation Theory) [Kamp 1988],
- La théorie des représentations discursives segmentées – SDRT (Segmented Discourse Representation Theory) [Lascarides 2007], et
- La technique de l'exploration contextuelle – EC [Minel 2009].

2.3.1 La grammaire de Montague

Selon Montague [Montague 1973], il n'y a pas de différence de principe entre la sémantique des langues naturelles et artificielles. Montague a utilisé des développements récents de la logique intentionnelle pour mettre en évidence la structure logique des langues naturelles.
Ainsi, cette méthode propre à *la grammaire de Montague* (GM) distingue d'une part entre le sens des expressions linguistiques, et d'autre part, la structure de l'ensemble des entités désignées (un modèle). En effet, la détermination de la signification d'une phrase P d'une langue revient à établir les conditions de vérité de P dans l'ensemble des mondes possibles. C'est en ce sens que nous parlons de sémantique vériconditionnelle.

La grammaire sémantique universelle de Montague s'oppose de manière critique à la théorie grammaticale de Chomsky [Chomsky 1959], dans laquelle la sémantique est considérée comme un composant indépendant de la syntaxe [Chomsky 2007]. Montague affirme au contraire que le sens d'une phrase est immédiatement lié à sa construction syntaxique [Gamut 1991].
Un des principes gouvernant la Grammaire de Montague (GM) est le principe de compositionnalité, du fait que, à chaque règle syntaxique correspond une règle sémantique. Le principe de compositionnalité chez Montague se base sur un concept mathématique représentant les structures syntaxiques, où l'espace des valeurs sémantiques comme des algèbres et l'interprétation sémantique comme un homomorphisme.
La GM se heurte à certains problèmes d'interprétation des pronoms au-delà des limites de la phrase, et en particulier, au problème de relations anaphoriques entre les pronoms et les descriptions définies. Pour cette raison, au début des années 80, certains travaux ont cherché des voies alternatives à l'approche montagovienne, parmi lesquelles on trouve la théorie des représentations discursives (DRT) [Kamp 1988].

2.3.2 La théorie de la structure rhétorique

Cette technique repose sur une analyse rhétorique, c'est-à-dire une vision linguistique de la structuration des textes, afin de détecter les relations sémantiques et les relations intentionnelles qui existent entre les segments d'un document. En effet, cette analyse rhétorique a comme but d'établir les relations et les dépendances ainsi que l'importance relative des phrases ou propositions les unes par rapport aux autres [Teufel 1998].
En 1988, Mann et Thompson [Mann 1988]ont réalisé une étude analytique et ils ont pu déduire, suite à leurs observations empiriques, qu'il est possible d'analyser la majorité des types des texte en termes d'arbre hiérarchique des relations rhétoriques (les relations sémantiques et les relations intentionnelles) qui existent entre les propositions d'un document.
Les résultats de cette étude ont permis à ces deux chercheurs de définir une théorie descriptive et fonctionnelle sur l'organisation des textes qu'ils ont nommé la RST (Théorie des Structures Rhétoriques).
Les auteurs de cette théorie posent une vingtaine de relations rhétoriques permettant de lier deux segments de texte adjacents entre eux, dont l'un possède le statut de noyau – segment de texte primordial pour la cohérence – et l'autre celui de satellite – segment optionnel [Luc 2001]. En effet, le but de cette théorie consiste en premier lieu à faire la distinction entre les deux types de segments : noyau, et satellite [Marcu 1997]. Les auteurs de la RST définissent les unités minimales comme des unités fonctionnellement indépendantes : elles correspondent généralement aux propositions [Luc 2001]. En deuxième lieu, le but de l'RST se concentre à faire la reconnaissance des relations qui relient le segment minimal avec celui composé : la reconnais-

sance repose sur une interprétation. Cette interprétation se base sur une analyse syntaxique du contenu des segments du texte afin de repérer les différents marqueurs lexicaux indicateurs de relations rhétoriques, c'est-à-dire décrire les relations qui lient deux parties du texte dont l'un est noyau et l'autre satellite. Ainsi, les relations peuvent être repérées via des signaux linguistiques. Des exemples de relations et de marques associées sont donnés dans le tableau suivant :

TABLEAU 2.2 – Exemples de relations rhétoriques et de marques associées [Ono 1996]

Relation	**Expressions**
Serial <SR>	Thus
Summarization <SM>	After all
Negative <NG>	But
Example <EG>	For example
Especial <ES>	Particularly
Reason <RS>	Because

Chaque relation est définie par des contraintes sur le noyau, sur le satellite et sur la combinaison de la paire noyau et satellite. Ainsi, pour déduire la relation entre deux propositions il faut appliquer des jugements de plausibilité (par exemple le lecteur ne croit pas N mais si on lui dit S alors il croit N avec une certaine certitude) [Saggion 2000].
L'autre élément qui s'ajoute à la théorie, c'est les schémas qui spécifient la composition structurale du texte. Ces schémas rhétoriques décrivant l'organisation structurelle d'un texte, quelque soit le niveau hiérarchique de ce dernier, permettent de lier un noyau et un satellite, deux ou plusieurs noyaux entre eux, et un noyau avec plusieurs satellites [Luc 2001].
Ainsi, les schémas rhétoriques se présentent sous la forme de cinq modèles de schémas (figure 2.3) qui peuvent être utilisés récursivement pour décrire des textes de taille arbitraire. En plus, un schéma indique comment co-occurrent les blocs textuels arguments d'une relation rhétorique.

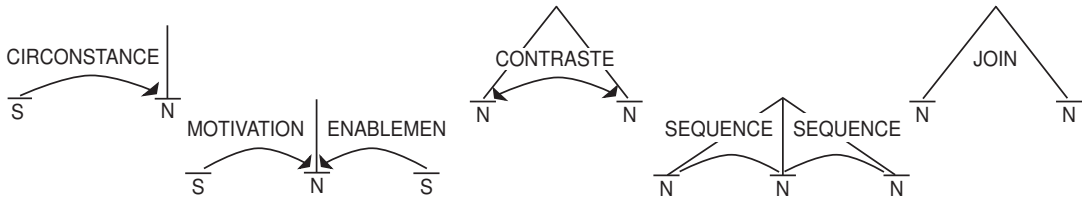

FIGURE 2.3 – Cinq modèles de schémas rhétoriques [Mann 1988]

Le schéma le plus fréquent est celui liant un satellite unique à un noyau unique [Saggion 2000]. La structure du texte est donc définie en termes de compositions d'applications de schémas, et ce de manière réitérative. La structure rhétorique finale d'un texte est strictement hiérarchique et se présente sous la forme d'un arbre RST.
Pour analyser un texte selon la RST, il doit d'abord être segmenté en unités textuelles. En général, ce sont des propositions et représentent les nœud terminaux de l'arbre RST qui est obtenu à partir de l'application d'un certain nombre de schémas. Dans la Figure 2.4, on présente une interprétation RST des textes suivants.
Exemple 1 :

(أ) تشتهر مدينة صفَاقس بتقديم أطبَاق ثمَار البحر علَى أنوَاعهَا. (ب) صفَاقس مدينة سَاحلية.

(O) t$thr mdynp SfAqs btqdym OTbAq vmAr AlbHr ElY OnwAEhA. (b) SfAqs mdynp sAHlyp.
(A) La ville de Sfax est connue par la présentation des plats de fruits de mer de tout type. (B) Sfax est une ville littorale.

La RST va réagir avec cet exemple comme suit (Figure 2.4) :

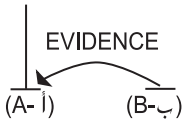

FIGURE 2.4 – Arbres RST de l'exemple 1

Exemple 2 :

(أ) تشتهر مدينة صفَاقس بتقديم أطبَاق ثمَار البحر علَى أنوَاعهَا. (ب) عندمَا يرتَاد زوَار مدينة صفَاقس (ت) يطلبون بَاستمرَار أطبَاق ثمَار البحر وخَاصة طبق المحَار وَالأَخطبوط المشوي علَى الفحم.

(O) t$thr mdynp SfAqs btqdym OTbAq vmAr AlbHr ElY OnwAEhA. (b) EndmA yrtAd zwAr mdynp SfAqs (t) yTlbwn bAstmrAr OTbAq vmAr AlbHr wxASp Tbq AlmHAr wAlIxTbwT Alm$wy ElY AlfHm.
(A) La ville de Sfax est connue par la présentation des plats de fruits de mer de tout type. (B) Lorsque les visiteurs se rendent à la ville de Sfax, (C) ils demandent régulièrement les plats de fruits de mer et surtout le plat d'huître et de poulpe grillé sur le charbon.

La RST va se réagir avec cet exemple comme suit (Figure 2.5) :

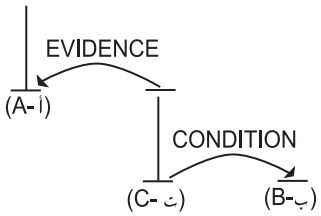

FIGURE 2.5 – Arbres RST de l'exemple 2

Dans ces approches, l'objectif du processus d'interprétation est d'obtenir l'arbre rhétorique du texte original.
Selon Marcu [Marcu 2000b], et suite à ces études empiriques, le meilleur arbre RST le plus descriptif est celui le plus équilibré à droite et à gauche.
Ainsi, ces propos offrent un milieu d'investigation assez intéressant pour le domaine d'extraction automatique. En effet, lors de l'étape de sélection des unités importantes, on peut profiter des relations entre les structures de discours pour décider le degré de leur importance. L'étape de réduction peut éventuellement exploiter les relations de cohérence et de cohésion pour garder l'unité textuelle de l'extrait et pour éviter les ruptures de séquences.
Les expériences avec l'utilisation de RST montrent que l'arbre rhétorique permet de prédire assez bien les unités qu'un juge humain aurait sélectionnées. Mais tel qu'a affirmé 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).

2.3.3 La théorie de la représentation discursive

La Théorie des Représentation Discursives (Discoure Representation Theory : DRT), proposée par Hans Kamp [Kamp 1981], [Kamp 1988], [Kamp 1993] et comme d'autres théories du discours se base sur une sémantique formelle. Cette représentation formelle a pour objectif de déterminer les conditions de vérité d'un segment et d'interpréter sémantiquement le discours (suites cohérentes de phrases ou segments). Ainsi, l'objectif de la théorie est de représenter un énoncé sous une représentation structurelle [Crabbé 2007].
Cette théorie a la particularité de postuler un niveau intermédiaire de représentation entre la syntaxe des phrases et leur interprétation dans un modèle [Gamut 1991]. Les structures de ce niveau intermédiaire sont appelées Structures de Représentation Discursive (Discourse Reprensentation Structures – DRS). La théorie DRT contient en conséquence deux composant principaux :

1. La procédure de construction des DRS, qui est le mécanisme qui permet de passer d'un énoncé ou d'un ensemble d'énoncés à cette représentation intermédiaire, et

2. La méthode au moyen de laquelle une DRS est interprétée dans un modèle, ce que Nicholas Asher appelle « définition de correction » [Asher 1993].

Les évolutions de la DRT ont été menées en intégrant les événements et leurs relations dans la construction du discours. Ainsi, ces évolutions sont remarquées au niveau de la construction et de la correction des DRS [Asher 2001].
Cette théorie a fait l'objet de plusieurs axes de recherches, en particulier ceux liés à la résolution de l'anaphore [Corblin 2001], la construction des structures de représentation [Amsili 2004], le traitement du temps [Reboul 1996] [Reboul 1998a] [Reboul 1998b] [Moeschler 2011], la prise en compte du pluriel [Estratat 2004], etc.
Dans ce qui suit, nous donnons une description des deux composantes principales de la théorie DRT à savoir l'interprétation de la DRS et sa correction.

Représentation de DRS Pour commencer, nous présentons un exemple de Structure de Représentation Discursive - DRS représentant une phrase simple 2.3. Cette phrase est traduite

TABLEAU 2.3 – Exemple d'énoncé avec sa représentation formelle

Énoncé	**Représentation formelle de l'énoncé**
Michel mange une tarte.	**K** = <x,y, Michel(x), tarte(y), x mange y>

par une DRSK relative à la représentation formelle K, et représentée par la suite graphiquement en « boîte » 2.6. Ainsi, comme nous montre l'exemple précédent, la DRS_k est une structure

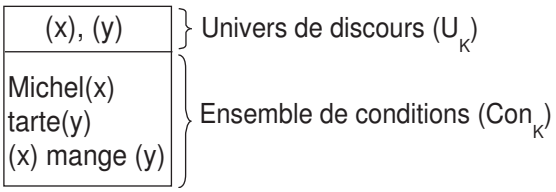

FIGURE 2.6 – La représentation de la DRS_K

composée de deux ensembles [Amblard 2007] [Amblard 2008].
- Le premier, {x,y} qui regroupe des éléments appelés référents de discours et que l'on peut comparer à des variables en logique du premier ordre. Ils sont destinés à être liés (mis en correspondance) par une fonction d'assignation à des individus du modèle. Cet ensemble est appelé univers de discours de laDRS, noté U_k.
- Le second ensemble, {Michel(x), tarte(y), x mange y} qui regroupe des conditions portant sur les référents de discours de la DRS. Ces conditions peuvent être :

1. Simples, c'est-à-dire des prédicats appliqués à des référents de discours (x et y dans notre exemple), ou
2. Complexes, c'est-à-dire mettant en jeu d'autres référents relatifs à d'autres DRS (dites subordonnées).

L'objectif des conditions (simples ou complexes) est analogue à celui des prédicats de la logique du premier ordre. Ce second ensemble est appelé ensemble de conditions, noté Con_k.
Une DRS_k est donc un couple $<U_K, Con_K>$, où :

- U_k est un ensemble de référents de discours (univers de discours), et
- Con_k est un ensemble de condition

Correction de la DRS La définition de correction de la DRS se fait par l'intermédiaire des règles associées aux catégories syntaxiques des mots du discours. La nécessité de cette étape réside dans le fait qu'elle essaie au maximum d'aller à un niveau d'analyse discursif plus détaillé. Par exemple, pour résoudre le problème d'anaphore et selon Corblin, [Corblin 2001], le nom propre a un traitement particulier, il lui associe un marqueur et une condition d'égalité qui le caractérise. Ce nom propre sera ensuite remplacé par son marqueur dans le discours.
En DRT, la même DRS croît au fur et à mesure de l'analyse du discours, car une DRS représente un texte de plusieurs propositions. La théorie de la représentation discursive segmentée SDRT n'utilise pour sa part que des DRS élémentaires (limitées à une proposition élémentaire), reliées par des relations de discours.

2.3.4 Techniques de la théorie de la représentation discursive segmentée (SDRT)

Cette théorie est initialement présentée par Acher et Lascarides [Asher 1993], [Lascarides 1993]. Ces deux chercheurs ont exploité l'analyse rhétorique du texte donnée par la Théorie de la structure rhétorique et la représentation formelle donnée par la Théorie des Représentation Discursives.
La théorie SDRT propose une représentation dynamique du discours qui tient en compte de la segmentation et de l'organisation structurelle discursive.
L'amélioration de cette théorie réside dans l'interaction entre le contenu sémantique des segments et la structure globale du discours [Lascarides 2009].
Comme la DRT, la SDRT est une théorie opératoire, du fait qu'elle vise à décrire une méthode déterministe de construction des Structures de Représentation du Discours Segmentées ou SDRS. Comme les DRS, les SDRS visent le contenu propositionnel du discours afin d'assurer une représentation macro-structure du discours.
Pour SDRT, les unités minimales sont les propositions et les relations de discours sont de

nature sémantique plutôt qu'intentionnelles. Ainsi, la SDRT adopte une méthode ascendante lors de la construction des représentations.
Toutefois et comme c'est le cas pour les DRS qui représentent les propositions, les SDRS qui représentent les segments complexes, au sens où l'on construit un segment complexe (composé de plusieurs segments élémentaires ou complexes) à partir d'autres segments. Dans ce processus de construction, tout nouveau segment (DRS ou SDRS) doit être relié à un segment précédent par une ou plusieurs relations de discours. En fait, la SDRT cherche ainsi, à décrire de façon systématique, dans un cadre logique, les mécanismes qui permettent aux locuteurs d'inférer les relations entre segments [Busquets 2001].
La représentation graphique associée à la SDRS est une "boîte" à deux compartiments, l'un pour les référents de discours appelés "étiquettes" et l'autre pour les conditions de SDRS .
Les SDRS les plus simples sont celles des discours à une seule proposition.
La SDRT reprend en plus des SDRS, les notions de cohérence introduisent par l'analyse du discours utilisée par d'autres théories comme la théorie RST.
De même, la SDRS a aussi une structure hiérarchique déterminée par le type de relation co-ordonnante ou subordonnante, reliant les différents segments entre eux.
En ce qui concerne l'attachement d'un nouveau segment à la structure discursive déjà construite, ce segment ne peut être attaché qu'au dernier segment analysé ou aux segments qui dominent hiérarchiquement ce dernier segment.
La SDRT abandonne la contrainte, adoptée par la RST [Mann 1988], qu'une seule relation de discours relie deux segments. En effet, plusieurs relations peuvent simultanément relier deux segments d'une SDRS [Asher 2001].

Exemple : Jean a donné un livre à Marie, mais il le lui a ensuite repris.

Cet exemple présente deux relations de discours entre ces deux propositions : *Contraste* et *Narration* ; indiquées par les marqueurs linguistiques "*mais*" et "*ensuite*".
Dans le cadre de la RST, le choix d'une relation ne dépend pas uniquement d'une propriété linguistique du texte ou des segments impliqués, mais surtout des intuitions de l'allocutaire par rapport aux intentions du locuteur ou auteur. C'est la raison pour laquelle Moore et Pollack [Moore 1992] ont soulevés le problème de la RST vis-à-vis des relations sémantiques et des relations intentionnelles. En cas d'ambiguïté (lorsque plusieurs relations sont vérifiées), l'allocutaire choisira une de ces relations en fonction des effets qui leur sont associés.
Si de multiples relations de discours peuvent réaliser un même attachement, la théorie doit expliquer non seulement comment on infère ces relations de discours mais aussi quelles sont les relations compatibles ou incompatibles [Asher 2001].
Une SDRS est donc à la base composée de DRS élémentaires, mais lorsqu'on est en présence

d'un discours structuré hiérarchiquement, sa représentation doit l'être aussi. On distingue alors dans la SDRS des éléments qui correspondent aux segments complexes du discours et qui sont eux-mêmes des SDRS : une SDRS est une structure récursive. L'ensemble de ces éléments, DRS et sous-SDRS, sont appelés constituants d'une SDRS.
Dans une SDRS, un énoncé est représenté par une formule du type $p : K$ et des relations rhétorique de type $R(p1, p2)$.
Une SDRS est un couple $< U, C >$, où :
- U est un ensemble d'étiquettes
- C est un ensemble de conditions
- p est une étiquette
- K le constituant (DRS ou SDRS) que désigne l'étiquette p
- R est une relation de discours des étiquettes $p1$ et $p2$

Afin de construire le graphe SDRT, la détermination de la structure hiérarchique dépend des relations définies entre les étiquettes des SDRS. La détermination de ces relations a été fréquemment liée à l'étude spécifique du corpus. De ce fait, il n'existe pas pour l'instant de liste définitive des relations de discours indépendantes du type de texte. Toutefois, nous pouvons indiquer les relations les plus fréquemment utilisées pour les textes narratifs [Asher 2001] : Narration, Arrière-Plan, Pré-condition, Commentaire, Élaboration, Topique, Continuation, Résultat, Explication, Parallèle, Contraste et Conséquence.
Les deux exemples suivants 2.7 présentent une SDRS hiérarchisée relative à la représentation formelle complexe :

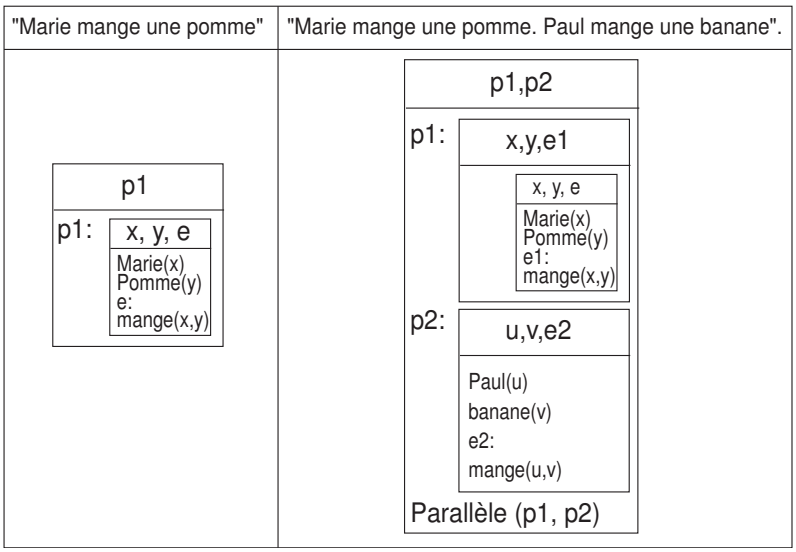

FIGURE 2.7 – Représentation de SDRS hiérarchisée [Busquets 2001]

Ainsi, comme nous avons indiqué précédemment, lorsqu'une condition $p : K_p$ porte sur un

constituant K_p qui n'est pas une simple DRS mais une SDRS complexe, on est en présence d'une SDRS hiérarchisée.
Le graphe SDRT final, est formé par des nœuds (constituants étiquetés) et les arcs où les attachements, ornés par les relations de discours.
La SDRT fait en effet l'hypothèse que dans tout discours cohérent, chaque constituant (sauf le premier) est attaché par une relation de discours à un constituant précédent. Plusieurs relations peuvent correspondre à un même attachement. La convention, en SDRT comme dans d'autres théories du discours, veut que l'arc soit vertical (de haut en bas) si la relation est subordonnante, et horizontal (de gauche à droite) si elle est coordonnante.
L'exemple suivant illustre le rôle de la hiérarchisation d'une structure SDRS complexe.

a. Jean est entré hier à l'hôpital.

b. Marie lui a cassé le nez,

c. et Paul lui a cassé le bras.

d. Il a été opéré tout de suite

d'. Il l'a même mordu.

d". Elle l'a même mordu.

La représentation et le graphe de la SDRS sont les suivants 2.8 :

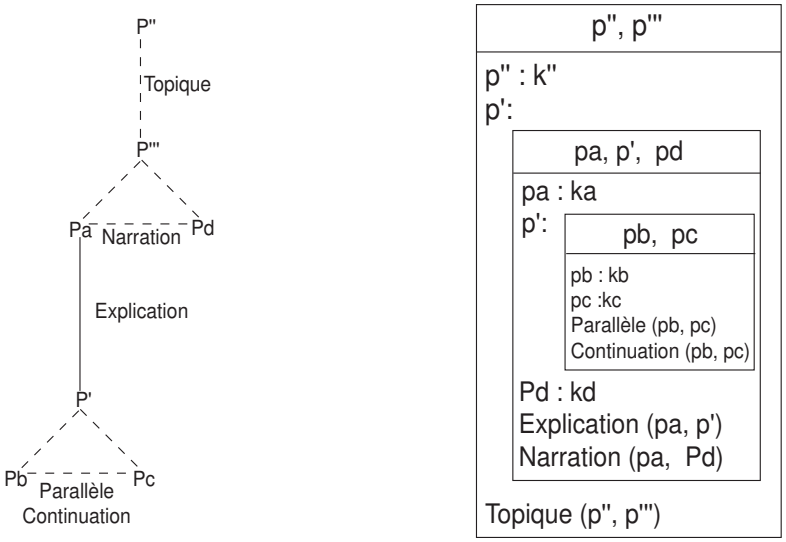

FIGURE 2.8 – Représentation et graphe de SDRS [Busquets 2001]

Dans ce graphe, les arcs présentés par des traits continus représentent les attachements entre étiquettes, ornés par les relations correspondantes. Les traits discontinus correspondent à la hiérarchisation entre SDRS et sous-SDRS. Ici, cela signifie que *pb* et *pc* sont des étiquettes de l'univers de discours de la sous-SDRS p', ce qui explique leur position la plus en-

châssée dans la SDRS car p', tout comme pa et pd, appartient à l'univers de p'''. Lorsqu'une relation a pour second argument une SDRS complexe, la relation se "distribue" sur son univers de discours : la sémantique de Explication (pa, p') est ici équivalente à la sémantique de Explication$(pa, pb)^E xplication(pa, pc)$. Ceci est une manifestation du principe Poursuite du Schéma Discursif (PSD, Continuing Discourse Patterns en anglais) qui garantit que tous les constituants d'un segment complexe ont un rôle homogène vis-à-vis de la partie supérieure de la structure.
Dans $(a - d)$, $(b - c)$ constitue un segment de discours qui explique pourquoi Jean est entré à l'hôpital. La relation Explication est subordonnante, ce qui correspond à l'observation linguistique que l'on peut poursuivre une explication $((c)continue(b))$, pour revenir ensuite au niveau de l'énoncé expliqué avec (d), alors que l'inverse n'est pas vrai : le discours $(a - d)$ ne peut être prolongé par (d'). La SDRT rend compte du fait que (d') peut prolonger $(a - c)$, mais pas $(a - d)$, par la notion de sites disponibles pour l'attachement, notion elle-même basée sur celle de frontière droite. La détermination des référents disponibles pour la résolution des anaphores repose également sur la notion de subordination dans la structure.
Considérons par exemple la SDRS du discours $(a - c, d')$, c'est-à-dire, le sous graphe du graphe ci-dessus n'incluant que pa, p',pb et pc, dans lequel on attachera (par Continuation) à pc le constituant étiqueté par pd représentant (d'), puisque (d')poursuit l'explication de (a). Les anaphores des deux pronoms $il et l'$de (d') peuvent être résolues, la première par une coréférence à un référent de discours introduit dans Ka (Jean), et la seconde par un référent introduit dansKc (Paul). En effet, ces référents sont tous les deux disponibles depuis pd, attaché à pc. Le cas est différent pour $(a - c, d'')$. La SDRT prédit que ce discours n'est pas bon par le fait que le pronom elle ne peut être résolue, les référents de Kb n'étant pas disponibles.

2.3.5 Techniques de l'exploration contextuelle

La méthode de classification en effectuant une exploration contextuelle [Desclés 1997], [Abraham 1992], [Desclés 2006], [Djioua 2007], [Le Priol 2009] s'est progressivement développée par une série de recherches qui ont évolué à travers les années [Berri 1996], [Aliod 1998], [Minel 2002b], [Crispino 2003], etc. Dans cette méthode, les phrases sont classifiées en catégories sémantiques hiérarchisées (Hypothèse, Objectif, Définition, Soulignement, etc.). Pour en déduire la catégorie sémantique, des règles d'exploration contextuelle doivent être appliquées à la phrase. Les règles d'exploration contextuelle attribuent une étiquette sémantique à une phrase lorsqu'un ensemble de marqueurs co-occurrent. En fait, elles constituent un ensemble de stratégies décisionnelles qui permettent de construire des représentations à l'aide d'un examen des éléments linguistiques à l'intérieur de leur contexte textuel. Les stratégies décisionnelles de l'exploration contextuelle identifient en premier lieu un indicateur pertinent caractéristique du

problème à résoudre. Ensuite, le contexte linguistique est fouillé pour rechercher des indices linguistiques co-présents qui permettent de prendre la décision adéquate.
La forme des règles est la suivante (voir algorithme 2) :

Algorithme 2 La forme des règles d'exploration contextuelle

1: **soit** le contexte C
2: **if** (l'on repère un indicateur pertinent ILP dans C) **et** (l'on constate la présence d'indices linguistiques ILTi dans C) **et** (les contraintes CcK sont vérifiées) **then**
prendre la décision Di

Une instanciation de ce patron type est par exemple :

Algorithme 3 Exemple de la règle d'exploration contextuelle : RH3

1: **soit** le contexte C : une phrase
2: **if** (l'on repère un indicateur pertinent ILP appartenant aux classes L6 ou L25 dans C) **et** (l'on constate la présence d'indices linguistiques ILT1 appartenant à la classe L27-2 ou ILT2 appartenant à la classe L32 dans C) **et** (les contraintes CcK sont vérifiées : ILP et ILTi appartiennent à la même proposition) **then**
prendre la décision Di : La phrase est un énoncé d'hypothèse

Avec :
- C = énoncé courant (une phrase)
- L6 = je, nos, nous
- L25 = il doit être possible, il est possible que, ...
- L27-2 = attendre, prédire, supposer, ...
- L32 = confirmer, conséquence, effet, infirmer, ...

Cette règle 3 permet de déterminer que la phrase "*j'affirme que nous pouvons supposer que ...*" est une spécification d'hypothèse. Remarquons que la phrase du type "*j'affirme que nous ne pouvons pas supposer que ...*" semble aussi pouvoir déclencher l'exécution de la règle à moins que la négation soit traitée par d'autres règles.
L'implémentation logicielle de cette méthode de classification par exploration contextuelle a été concrétisée initialement par le système SERAPHIN [Berri 1996] qui a intégré au cours des années, de nouvelles fonctionnalités sous la forme de modules supplémentaires nommés SAFIR [Berri 1996], ContextO [Minel 2002a] et EXCOM-2 [Alrahabi 2010].

2.3.6 Discussion des méthodes symboliques

Nous avons essayé de monter ci-dessus, les principales méthodes reposant sur l'analyse du discours et de sa structure. Ces méthodes symboliques se basent, essentiellement, sur la détec-

tion de marques linguistiques, placent quant à elles l'importance d'une phrase dans la reconnaissance d'une marque particulière. Suivant le cas, cette marque est révélatrice, ou non, d'une catégorie dans laquelle peut être rangée la phrase (rôle, fonction, nature, position, etc.). On peut admettre, a priori, que l'on puisse trouver un ensemble de marques efficaces pour chaque type ou genre de texte. Cela n'est sûrement pas aussi simple pour certains genres littéraires ou lorsque le sous-entendu et les pré-requis dominent l'argumentation logique explicite. Ces méthodes permettraient néanmoins le traitement de nombreux genres de texte et de nombreuses langues. Par ailleurs, elles permettent le contrôle du contenu du résumé et donc l'adaptation aux besoins du lecteur. Ces deux aspects répondent bien à quelques-uns de nos objectifs ciblés pour le résumé automatique.

Les expériences rapportées sur la construction d'une représentation rhétorique du texte, qui ne sont d'ailleurs qu'une manière particulière d'exploiter des marques de surface, ont montré que l'arbre rhétorique permet de prédire assez bien les unités qu'un juge humain aurait sélectionné. Cependant, les relations rhétoriques seules, ne permettent pas de contrôler le contenu informationnel du résumé. 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. Les textes traités sont assez « courts » et il faudra se demander si l'approche continue d'être applicable, quand on passe du paragraphe au texte multi-paragraphe. La validité de l'approche n'est pas certaine non plus pour tous les genres textuels. En effet, dans le domaine technique, les écrivains marquent les relations entre les éléments explicitement, tandis que dans d'autres domaines, les relations sont plutôt sémantiques. Le cas des méthodes, reposant sur une représentation de type conceptuel avec utilisation de connaissances autres que linguistiques, est délicat à traiter. En effet, bien que ce genre de méthodes soit potentiellement le meilleur, il s'avère être aussi le plus décevant sur le plan de l'envergure des résultats produits. L'applicabilité de ces méthodes reste liée à un domaine particulier nécessitant des représentations et des connaissances spécifiques à ce domaine. Ces méthodes sont utilisées pour traiter des textes courts comme les dépêches d'agence de presse ou les brèves documentations techniques. Ainsi, ce qui est en cause dans ce type de méthodes n'est pas tant la démarche utilisée ou le modèle choisi mais plutôt les connaissances à "injecter" dans le système pour qu'il soit opérationnel. A défaut de processus d'acquisition automatique de connaissances, il paraît difficile à court terme de construire un système basé sur ce type de méthodes fonctionnant en domaine ouvert. Par ailleurs, la notion d'importance est considérée du point de vue de l'auteur, alors que selon nos objectifs le point de vue du lecteur du résumé doit être considéré en parallèle.

En résumé, comme nous pouvons le voir, aucune méthode ne semble réellement satisfaisante. Les méthodes statistiques sont aisées à mettre en œuvre mais ne fournissent pas de bons résultats, les méthodes linguistiques semblent possibles à mettre en œuvre mais les genres de texte

traités sont déjà plus limités. Quant aux méthodes conceptuelles, leur faisabilité est problématique jusqu'à l'heure actuelle.
Nous pouvons alors nous poser la question de la combinaison de plusieurs méthodes. C'est ce que différents chercheurs ont tenté (voir précédemment) en mariant par exemple l'analyse de fréquence et la détection de motifs lexicaux. Ce type d'approche hybride nécessite de contrôler finement la part de chaque technique dans le résultat final. Autrement dit, il faut être capable de décider quelle technique privilégier en fonction des données.

2.4 Approche hybride

Étant donné que les résumés produits sont généralement peu cohérents à cause de l'extraction de phrases déconnectées de leur contexte, les approches hybrides essayent de combler cette lacune en proposant des méthodes numériques qui tiennent en compte les traits du discours qui assurent sa compréhension.
La majorité des travaux sur l'extraction sont fondé sur l'extraction basée sur les connaissances hybrides provenant de différentes sources symboliques et numériques ou même à base d'attribution de poids/score à des informations extraite d'une façon symbolique [Ono 1994], [Marcu 2000a]. Dans ce contexte d'idée, Ono et Marcu font précéder l'étape d'extraction par une présentation du texte source sous forme d'un arbre RST tout en l'assignant un poids a ses différents nœuds en fonction de leur position. En effet, c'est le sore final qui juge la pertinence des nœuds d'un arbre. Ainsi, la sélection de phrases pour le résumé est effectuée en fonction de la longueur désirée du résumé, et le choix est plus ou moins d'éléments, sera fixé selon dans un ordre déterminé par l'algorithme de sélection [Torres-Moreno 2011].
Après étude des différentes approches pour le domaine du résumé automatique, ainsi que les différentes méthodes utilisés, a notre connaissance, il n'existe pas des travaux de recherche qui ont résolu le problème d'extraction à base d'un contrôle terminant la part de chaque technique dans le résultat final. Il faut, cependant, noter que le traitement du problème de résumé en combinant les méthodes numériques et symboliques, pourrait permettre de franchir un palier et de s'approcher un peu plus de ce que peut faire les résumeurs humains. Le paradigme d'extraction des phrases en se basant sur une approche hybride qui privilégiera l'utilisation des techniques numériques et symboliques en fonction des données peut servir, à notre point de vue, à être un pas en avant vers la génération d'extrait d'une meilleure qualité.

2.5 Conclusion

Dans ce chapitre, nous avons exploré quelques méthodes qui ont été proposées pour résoudre le problème de production de résumés automatiques. Il semble bien que les méthodes purement

statistiques, simplement implantées et rapidement adaptables à d'autres domaines soient limitées en ce sens qu'elles n'ont pas une vision globale du texte. Les méthodes fondées sur la compréhension nécessitent des modèles conceptuels et des ressources linguistiques avancées et ne peuvent être donc appliquées qu'à des domaines restreints.
En revanche, les méthodes hybrides qui tiennent compte à la fois des techniques symboliques et numériques sont plus prometteuses. Malgré les avancées enregistrées par ces méthodes, la lisibilité des résumés reste toujours à améliorer.
À partir de cette synthèse des méthodes proposées pour le résumé automatique, nous pouvons conclure que pour le développement de systèmes de résumés automatiques, une approche exploratoire du document source est plus avantageuse. Étant donné que les résumés produits sont généralement peu cohérents à cause de l'extraction de phrases déconnectées de leur contexte, nous allons essayer de combler cette lacune en proposant une approche hybride basée sur des relations rhétoriques de résumés conçus à partir d'une vision globale du document source.

Deuxième partie

Les bases théoriques et techniques pour une nouvelle approche

Chapitre 3

Les techniques de TALN pour le résumé automatique de l'arabe

Sommaire

3.1 Introduction 47
3.2 Particularités de la langue arabe 48
3.2.1 Absence de voyelles 49
3.2.2 Agglutination 50
3.2.3 Irrégularité de l'ordre des mots dans la phrase 50
3.2.4 Absence de ponctuation régulière 51
3.3 Difficultés de l'analyse automatique de l'arabe 51
3.3.1 La segmentation de textes 51
3.3.2 L'analyse morphologique 51
3.3.3 L'étiquetage grammatical 52
3.3.4 L'analyse syntaxique 53
3.4 Principales approches de traitement automatique de l'arabe écrit . . 54
3.4.1 Approches de Segmentation de textes 54
3.4.2 Approches d'analyse syntaxique 56
3.4.3 Approches d'étiquetage grammatical 58
3.4.4 Approches d'analyse morphologique 58
3.4.5 Approches de reconnaissance des entités nommées 59
3.5 Conclusion 61

3.1 Introduction

L'arabe est parlé dans plus de 22 pays, du Maroc jusqu'à l'Iraq et dans toute la péninsule arabe [Versteegh 2001]. C'est la première langue pour plus de 250 millions de personnes et la deuxième pour 40 millions. L'arabe, langue du Coran, est devenue la langue d'une civilisation et ne sert plus seulement à désigner les seuls habitants de la péninsule arabe qui la parlaient.

On distingue l'arabe classique et l'arabe moderne. L'arabe classique est la forme littéraire utilisée par tous pour les besoins de l'écriture et de l'imprimerie. C'est aussi la langue de la religion pour tous les musulmans, quelle que soit par ailleurs leur langue vernaculaire. L'arabe moderne, dérivé de l'arabe classique, est la langue de la presse, des débats politiques, des textes scientifiques et de plus en plus celle des textes littéraires profanes. Parlé dans la plupart des pays arabes, l'arabe moderne n'est en revanche presque jamais la langue des échanges quotidiens.
Depuis le début des travaux en Traitement Automatique du Langage Naturelle sur la langue arabe, plusieurs études ont poursuivi des directions de recherche diverses. On peut notamment distinguer les approches numériques s'appuyant sur des probabilités et les statistiques et des approches symboliques liées à la théorie des langages formels. Comme nous le verrons dans ce chapitre, ces études vont de l'analyse détaillée de phrases isolées à des approches plus globales d'un texte dans son ensemble.
Dans ce chapitre, nous présentons d'abord les particularités de la langue arabe et les principales ambiguïtés rencontrées lors de son analyse. Ensuite, nous donnons un aperçu sur les approches d'analyse de l'arabe se rapportant aux niveaux lexical, syntaxique et sémantique. Enfin, nous présentons les principales contributions effectuées dans le domaine du TALN arabe écrit.

3.2 Particularités de la langue arabe

Vue ses propriétés morphologiques et syntaxiques, la langue arabe est considérée comme une langue difficile à maîtriser dans le domaine du TALN.
Les premiers travaux de recherche, débutés vers les années soixante-dix, ont concernés par les lexiques et la morphologie arabe.
Avec l'avènement de l'Internet et des moteurs de recherche, la quantité de documents arabes disponibles en format électronique est devenue énorme. De ce fait, plusieurs travaux de recherche pour le traitement automatique de l'arabe commencent à émerger.
Ces travaux ont pris diverses orientations de recherche se rapportant à la syntaxe, la traduction automatique, l'indexation automatique de documents, la recherche d'information, etc.
Ces travaux sur le traitement automatique de l'arabe ont toujours fait face aux problèmes variés de la langue arabe issus de la nature agglutinante de cette langue, sa richesse flexionnelle, l'absence de vocalisation de la majorité des textes arabes écrits, etc.
Dans la section suivante, nous essayerons de donner une brève présentation de ces problèmes, qui rendent le traitement automatique de la langue arabe une tâche difficile à maîtriser.

3.2.1 Absence de voyelles

Le problème de la voyellation réside dans l'absence quasi systématique de la voyellation dans les textes arabes. En effet, les signes de voyellation, lorsqu'ils sont notés, sous la forme de signes diacritiques placés au dessus ou au dessous des lettres, apparaissent dans certains textes (Coran, hadith) ou littéraires (poésie classique, notamment) : on dit qu'ils sont édités en graphie voyellée.

L'absence de voyelles (la non-voyellation) dans les textes arabes génère plusieurs cas d'ambiguïtés et des problèmes lors de l'analyse automatique. En effet, l'ambiguïté grammaticale augmente si le mot est non voyellé. Cela est dû au fait qu'un mot non voyellé possède plusieurs voyellations possibles, et pour chaque voyellation est associée une liste différente de catégories grammaticales [Belguith 1999].

L'exemple suivant 3.1 du mot non-voyellé ktb | كتب possède 16 voyellations potentielles et qui représentent 9 catégories grammaticales différentes [Debili 2002].

Tableau 3.1 – Exemple de voyellation [Debili 2002]

Mot voyellé	Pré-notion	Notion d'écrire
كَتَبَ	kataba	Il a écrit
كُتِبَ	kutiba	Il a été écrit
كُتب	kutub	Des livres
كَتب	katob	Un écrit
كَتَّبَ	kattaba	Il a fait écrire
كُتِّبَ	kuttiba	Faire écrire- forme factitive
كَتِّبُ	kattibo	Fais écrire
كَتَبَّ	katabba	Comme trancher
كَتَب	katabb	Comme 'tranchement'
...	...	...

3.2.2 Agglutination

Contrairement aux langues latines, en arabe, les articles [1], certaines prépositions [2], les pronoms [3], etc. collent aux adjectifs, noms, verbes et particules auxquels ils se rapportent. Comparé au français, un mot arabe peut parfois correspondre à une phrase française (Souissi, 1997).
Exemple : le mot arabe "أتـتذكرونـنَا" correspond en Français à la phrase "*Est-ce que vous vous souvenez de nous ?*".
Cette caractéristique peut engendrer une ambiguïté au niveau morphologique. En effet, il est parfois difficile de distinguer entre une proclitique [4] ou enclitique [5] et un caractère original du mot. Par exemple, le caractère "و" dans le mot "وصل" (*il est arrivé*) est un caractère original alors que dans le mot "وفتح" (il a ouvert), il s'agit d'une proclitique [Belguith 2006].

3.2.3 Irrégularité de l'ordre des mots dans la phrase

L'ordre des mots en arabe est relativement libre. D'une manière générale, on met au début de la phrase le mot sur lequel on veut attirer l'attention et l'on termine sur le terme le plus long ou le plus riche en sens ou en sonorité. Cet ordre provoque des ambiguïtés syntaxiques artificielles, dans la mesure où il faut prévoir dans la grammaire toutes les règles de combinaisons possibles d'inversion de l'ordre des mots dans la phrase [Belguith 2005].
Ainsi par exemple, on peut changer l'ordre des mots dans la phrase (voir 3.2) pour obtenir deux phrases ayant le même sens.

TABLEAU 3.2 – Exemple de combinaisons possibles d'inversion de l'ordre des mots dans la phrase [Belguith 2005]

Verbe + sujet + complément	فعل + فَاعل + متمم	Est allé le garçon à l'école	ذهب الولدُ إِلَى المدرسة
sujet + verbe + complément	فَاعل + فعل + متمم	Le garçon est allé à l'école	الولدُ ذهب إِلَى المدرسة
complément + verbe + sujet	متمم + فعل + فَاعل	A l'école est allé le garçon	إِلَى المدرسة ذهب الولدُ

1. Les articles : par exemple "الـ"
2. Les prépositions sont : في، مع، لن، عن، علَى، حتَّى، من، إِلَى، ل، ك، ب
3. Le pronom personnel en arabe est isolé ou affixé. Isolé, il correspond en français à : moi, toi, etc.
4. Les proclitiques représentent des conjonctions mono-consonnes (و، لـ), des prépositions (لـ، بـ), un préverbe (سـ) indiquant le futur, un article (الـ) qui permet la détermination d'un nom, etc.
5. Les enclitiques sont les compléments de pronom ك، كمَا، هم، ـه ، ...

3.2.4 Absence de ponctuation régulière

La langue arabe n'est pas basée principalement sur les signes de ponctuations et les marqueurs typographiques ; il est à noter que ces derniers ne sont pas utilisés de façon régulière dans les textes arabes actuels, et même dans le cas où ils y figurent, ils ne sont pas gérés par des règles précises d'utilisation. [Belguith 2008]
Par ailleurs, nous pouvons trouver tout un paragraphe arabe ne contenant aucun signe de ponctuation à part un point à la fin de ce paragraphe. Ainsi, il convient de noter que la présence des signes de ponctuation ne peut pas guider la segmentation comme c'est le cas pour d'autres langues latines, telles que le français ou l'anglais. Ainsi, la segmentation de textes arabes doit être guidée non seulement par les signes de ponctuations et les marqueurs typographiques mais aussi par des particules et certains mots tels que les conjonctions de coordination, etc. [Belguith 2008].

3.3 Difficultés de l'analyse automatique de l'arabe

3.3.1 La segmentation de textes

La segmentation d'un texte est une étape fondamentale pour son traitement automatique ; son rôle est de découper le texte en unités d'un certain type qu'on aura défini et repéré préalablement. En effet, l'opération de segmentation d'un texte consiste à délimiter les segments de ses éléments de base qui sont les caractères, en éléments constituants différents niveaux structurels tels que : paragraphe, phrase, syntagme, mot graphique, mot-forme, morphème, etc. [Mouelhi 2008].
Toutefois, les particularités de la langue arabe, rend la segmentation arabe toujours différente, il n'y a pas de majuscules qui marquent le début d'une nouvelle phrase. De plus, les signes de ponctuation, ne sont pas utilisés de façon régulière (voir section 3.2).
D'après l'étude réalisée par Belguith [Belguith 2005], certaines particules comme "et | و", "donc | ف", etc. jouent un rôle principal dans la séparation de phrases et peuvent être déterminantes pour guider la segmentation.

3.3.2 L'analyse morphologique

L'opération de l'analyse morphologique tient à étudier la forme d'un mot en faisant une analyse interne de la structure de ce dernier. Le but étant de décomposer un mot à des éléments plus petits (préfixes, suffixes, etc.) selon des règles de combinaison relatives à ces derniers.
À proprement parler, l'analyse morphologique ne fait que la séparation et l'identification des morphèmes semblables aux mots préfixés (comme les conjonctions "wa | و" et "fa | ف", etc.),

des prépositions préfixées (comme "bi | ب" et "li | ل", l'article défini"الـ", etc.), des suffixes de pronom possessif [Chaaben 2010].
La phase d'analyse morphologique détermine un schéma possible. Les préfixes et suffixes sont trouvés en enlevant progressivement des préfixes et des suffixes et en essayant de faire correspondre toutes les racines produites par un schème afin de retrouver la racine.
Le problème principal de cette analyse réside dans l'agglutination et l'absence de voyellation.
Pour l'agglutination et contrairement aux langues latines, en arabe, les pronoms, les prépositions, les articles, les conjonctions, et autres particules collent aux noms, verbes, adjectifs et particules auxquels ils se rapportent. Comparé au français, un mot arabe peut parfois correspondre à une phrase française [Debili 1998].
Cette caractéristique engendre une ambiguïté morphologique au cours de l'analyse. Ainsi, la reconnaissance des unités lexicales qui composent une unité morphologique n'est pas toujours facile à détecter. Le problème est de reconnaître que la bonne segmentation réside ainsi, dans la difficulté de distinction entre un proclitique ou enclitique et un caractère original du mot. Par exemple, le caractère "و" dans le mot "il est arrivé | وصل" est un caractère original alors que dans le mot "et il a ouvert | وفتح", il s'agit plutôt d'une proclitique [Belguith 2006].
L'absence de voyellation pose un autre problème important. En effet, les mots non voyellés engendrent beaucoup de cas ambigus au cours de l'analyse (e.g. le mot non voyellé "فصل" pris hors contexte peut être un verbe au passé conjugué à la troisième personne du singulier "il a licencié | فَصَلَ", ou un nom masculin singulier "chapitre/ saison | فَصلٌ", ou encore une concaténation de la conjonction de coordination "puis | فَـ" avec le verbe "صل" : impératif du verbe lier conjugué à la deuxième personne du singulier masculin).

3.3.3 L'étiquetage grammatical

L'étiquetage grammatical est l'opération qui consiste à attribuer à chacun des mots d'un texte la catégorie (non, verbe, adjectif, article défini, etc.) qui est la sienne dans le contexte où il apparaît [Debili 2002].
La difficulté de l'étiquetage grammatical s'amplifie lorsque les textes visés se présentent sous leur forme non pas voyellée, mais partiellement voyellée ou encore totalement non voyellée, ce qui correspond au cas le plus courant.
Dans ces conditions, le but général de l'étiquetage grammatical consiste à répondre à la question suivante : Comment associer au différents mots qui composent un texte l'étiquette qui leur convient, compte tenu du contexte où ils occurrent ? Ainsi, le problème des étiquettes grammaticales est souvent posé lorsque les textes désirés sont sous leurs formes non ou partiellement voyellées, plutôt qu'à leurs formes voyellées [Debili 1998].

Le problème de la voyellation [6] d'un mot est ainsi posé du fait que le choix de l'accentuation qui convient au mot est difficile et dépend essentiellement du contexte [Belguith 2009].

Le tableau 3.3 présente le problème d'ambiguïté grammaticale rencontrée lors de l'attribution catégorique d'un mot non voyéllé "ktb | كتب", qui admet au moins cinq étiquettes grammaticales qui sont les suivantes :

TABLEAU 3.3 – Exemple d'étiquettes grammaticales attribuées selon la voyellation [Debili 2002]

Exemple de voyellation	**Étiquettes grammaticales**
كُتُبٌ | kutubun : des livres	substantif, masculin, pluriel
كَتبٌ | katbun : un écrit	substantif, masculin, singulier
كَتَبَ | kataba : il à écrit	verbe, 3ème personne masculin, singulier de l'accompli actif
كُتِبَ | kutiba : il a été écrit	verbe, 3ème personne masculin, singulier de l'accompli passif
كَتِّبْ | kattib : fais écrire	verbe à l'impératif, 2ème personne masculin, singulier

3.3.4 L'analyse syntaxique

L'analyse syntaxique permet d'associer à un énoncé sa ou ses structures syntaxiques possibles, en identifiant ses différents constituants et les rôles que ces derniers entretiennent entre eux.
Toutefois, l'analyse syntaxique prend en entrée le résultat de l'analyse lexicale (éventuellement de l'étiquetage morpho-syntaxique) et fournit en sortie une structure hiérarchisée des groupements structurels et des relations fonctionnelles qui unissent les groupements.
Enfin, il est à signaler que les ambiguïtés vocaliques et grammaticales, relatives à la non voyellation des mots, pose des difficultés au niveau de l'analyse syntaxique. Ainsi, une phrase, en absence de la voyellation, peut être interprétée et traduite selon plusieurs interprétations qui sont toutes syntaxiquement correctes.

6. L'opération qui consiste à restituer les signes vocaliques des voyelles de chacun des mots d'une phrase ou d'un texte non ou partiellement voyellé.
Cette opération est analogue a celle d'accentuations des mots en langue française.

3.4 Principales approches de traitement automatique de l'arabe écrit

Dans cette section, nous présentons les principales approches d'analyse de l'arabe qui couvrent principalement les niveaux de segmentation, analyse morphologique, étiquetage grammatical, reconnaissance des entités nommées, analyse syntaxique et analyse sémantique.

3.4.1 Approches de Segmentation de textes

Pour la plupart des applications de traitement automatique des langues naturelles (l'analyse de texte, l'extraction d'information, le résumé automatique) la segmentation devient une phase importante pour repérer les segments contenant les informations recherchées.
Ainsi par exemple, commencer une analyse d'un texte sans le segmenter en phrases conduit à des résultats peu fiables ; de même, avoir un mauvais segmenteur conduit à accumuler les erreurs du traitement automatique du texte [Ghassan 2001].
Malgré l'importance de cette étape dans l'analyse automatique d'un texte et qui a fait l'objet de certains travaux de recherche pour les langues latines, nous remarquons que les applications de segmentations destinées à la langue arabe ne sont pas nombreuses. À notre connaissance il n'existe que le système STAr [7] [Baccour 2004], [Belguith 2005] en état fonctionnel.
Ce système accepte en entrée un texte arabe non voyellé, de type TXT et produit en sortie un texte segmenté en phrases et propositions sous forme d'un fichier de type XML.
L'approche utilisée par le système STAr est purement symbolique, et se base essentiellement sur une technique d'exploration contextuelle des signes de ponctuation, des mots connecteurs (e.g. "lakin | لكن", "laqad | لقد", "amma | أمّا") jouant le rôle de séparateurs de phrases, ainsi que celles de certaines particules telles que les conjonctions de coordination ("wa | و" et "fa | ف") [Belguith 2005].
L'exploration contextuelle repose sur une étude des indices linguistiques déclencheurs appelés indicateurs et des indices complémentaires associés à ces indicateurs et sur un ensemble de règles [Desclés 1997]. Ces règles ont la forme suivante :

Algorithme 4 Format des règles de segmentation

1: **soit** un marqueur déclencheur **X**
2: **if** (le contexte gauche de **X** est **G**) **et** **ou** (le contexte droit de **X** est **D**) **then**
prendre la décision **Y** (fin ou non fin d'un segment).

Ces règles sont classées en trois classes relatives aux trois types de marqueurs déclencheurs à

7. STAr : système de Segmentation de Textes Arabes

savoir les signes de ponctuation, les particules et les mots connecteurs [Belguith 2005].
Le tableau 3.4 et l'algorithme 5 suivants présentent un exemple d'une règle relative à la virgule [Belguith 2005].

TABLEAU 3.4 – Exemple d'une règle de segmentation relative à la virgule

Contexte gauche		Marqueur		Contexte droit	
	Verbe	Espace	،		وفي صبَاح

Algorithme 5 Exemple d'une règle de segmentation relative à la virgule

1: **soit** la *virgule* est suivie par un *espace*
2: **if** (l'*espace* est suivi d'un *verbe*) **et** **ou** (le contexte droit de la virgule commence par " ") **then**
la virgule *ne marque pas* la fin de la phrase.

C'est le cas par exemple de l'énoncé suivant :

و في صبَاح مشرق من أصبَاح الصّيف، مرّ بَابن عمّه أسمَاعيل.

w fy SbAH m$rq mn OSbAH AlS yf, mr bAbn Em h IsmAEyl.

L'évaluation du système STAr a été réalisée sur deux corpus différents (voir tableau 3.5).

TABLEAU 3.5 – Corpus d'évaluation de STAr [Belguith 2005]

Corpus	**Nombre de textes**	**Nombre de paragraphes**	**Nombre de mots**
Deux livres de lecture	144	991	403 431
Articles de journaux	60	510	38 062

Les mesures de rappel et de précision obtenues pour le premier corpus [8] sont meilleures que ceux trouvés pour le deuxième corpus (voir tableau 3.6). Ceci s'explique par le fait que les articles de journaux contiennent des erreurs typographiques (i.e. insertion d'un espace après la conjonction de coordination "wa | و", omission de la lettre "chadda | الشدّة", des constructions erronées, etc.) qui augmentent le taux d'erreur au niveau de la segmentation en mots, de

8. Deux livres de lecture : 4ème année primaire et 9ème année de base

l'identification de la catégorie grammaticale des mots et par conséquent le taux d'erreur au niveau de la segmentation en phrases augmente.

Cette évaluation a montré aussi, que certaines ambiguïtés de segmentation ne peuvent être

TABLEAU 3.6 – Résultat de l'évaluation de STAr [Belguith 2005]

Corpus	Rappel	Précision
Livres	88.26%	80.65%
Articles de journaux	75.81%	65.66%

levées qu'à l'aide d'informations morphologiques. De plus, certaines particules utilisées dans la segmentation peuvent à leur tour être ambiguës.

3.4.2 Approches d'analyse syntaxique

Comme nous avons déjà entamé dans 3.3.4, l'analyse syntaxique permet d'associer à une phrase (ou un énoncé) sa ou ses structures syntaxiques possibles, en identifiant ses différents constituants et les rôles que ces derniers entretiennent entre eux [Bourigault 2000].

La plupart des travaux sur l'analyse automatique de textes arabes se sont focalisés sur des analyses de bas niveaux. Dans cette section, nous présentons un tour d'horizon des travaux réalisés sur l'analyse syntaxique de l'arabe, même si certains d'entre eux ne représentent qu'une simple adaptation d'approches ou de systèmes, initialement conçus pour d'autres langues.

Othman et al. [Chiraz 2003] ont proposé un analyseur syntaxique qui se base sur une approche symbolique. Cet analyseur utilise une grammaire d'unification et un "*chart parser*" [9] implémenté avec Prolog. L'information dans le lexique est combinée avec les caractéristiques de la tête (nom ou groupement nominal) pour éliminer certains choix proposés par l'analyseur morphologique. Aucune information n'est fournie quant à la couverture de la grammaire ou à la performance de cet analyseur [Belguith 1999].

Une évaluation chiffrée de la performance est donnée pour les deux analyseurs statistiques entraînés sur le corpus Treebank :

i l'analyseur de Collins ([Collins 2005] implémenté par Bikel [Bikel 2004a] est entraîné sur le corpus arabe TreeBank-1 [10] et a atteint un taux de rappel de 75.4% et un taux de précision de 76% sur les phrases de 40 mots au maximum et respectivement des taux de 72.5% et 73.4% pour toutes les phrases du corpus

9. "Chart parsing" représente l'analyse syntaxique tabulaire qui a pour principe de stocker les résultats intermédiaires dans une table (un "chart" en anglais) afin de pouvoir les réutiliser par la suite sans avoir à les recalculer.

10. ATB1 : http://www.ircs.upenn.edu/arabic/

ii Kulick et al. ([Maamouri 2008], [Gabbard 2008], [Kulick 2009]) ont utilisé l'analyseur de Bikel sur une version révisée de ATB1 et ont obtenu des résultats comparables à celles de Bikel. Ils l'ont ensuite testé sur ATB3 et la performance initiale a légèrement diminuée. Un certain nombre d'améliorations successives ont permis à l'analyseur d'atteindre un taux de rappel de 78.14% et un taux de précision de 80.26% sur des phrases de 40 mots au maximum et respectivement des taux de 73.61% et 75.64% sur toutes les phrases. Les deux plus importantes améliorations ont été obtenues en modifiant le traitement des signes de ponctuation et en choisissant un "*tagset*" [11] qui préserve un peu plus d'informations que celles distribuées par les segments de l'ATB.

Klein et al. [Klein 2003] ont utilisé l'analyseur *Stanford Parser* [12] pour l'arabe, mais aucune information sur sa performance n'est publiée. L'idée est de calculer la structure arborescente des syntagmes et les modèles d'arbre de dépendances lexicales, ensuite de les évaluer séparément. Des améliorations significatives peuvent, ainsi, être atteintes sans inclure aucune information sur la dépendance lexicale et ce en ajoutant quelques annotations linguistiques aux modèles de la structure arborescente.

Enfin, Chiang et al. [Chiang 2006] ont utilisé les deux analyseurs de Bikel [Bikel 2004b] et de Chiang [Chiang 2000] pour analyser l'Arabe Levantin (AL) [13]. Ils ont proposé une approche qui consiste à traduire l'AL à l'Arabe Standard Moderne (ASM) et ensuite à lier la phrase en AL aux analyses correspondantes en ASM.

Notons que la traduction automatique est particulièrement difficile quand il n'y a aucune ressource disponible comme les textes parallèles ou les lexiques de transfert. Ainsi, Chiang et al. [Chiang 2006] se sont basés principalement sur un corpus annoté de l'arabe moderne standard [14] ainsi que sur un corpus annoté de l'arabe levantin et plus précisément celui du dialecte jordanien [15]. Ils ont construit un lexique comportant des paires AL/ASM de formes de mots. Ils ont aussi construit une grammaire synchrone ASM-Dialecte. Ils assument ainsi que chaque arbre dans la grammaire de l'arabe moderne standard extraite du MSA Treebank est aussi un arbre de l'arabe levantin vu la similarité syntaxique entre l'ASM et l'AL.

La plupart des travaux que nous venons de citer ont pour objectif de faire une analyse syntaxique totale qui concerne tous les constituants de la phrase analysée. Cependant, d'autres travaux se sont intéressés à l'analyse syntaxique partielle [16] qui ne considère qu'un nombre restreint de mots dans la phrase ou qui se réduit uniquement à l'analyse des syntagmes locaux de la phrase, tels que les groupes nominaux et les constructions de verbes, sans pour autant en calculer sa

11. i.e., une liste d'étiquettes
12. http://nlp.stanford.edu/downloads/parser.shtml-lex
13. Approches d'étiquetage grammatical
14. i.e., MSA Treebank [Maamouri 2008]
15. i.e., LATB Treebank [Maamouri 2008]
16. i.e., une analyse non détaillée

structure syntaxique totale. Cette analyse permet de construire une structuration partielle des phrases à analyser dans un but précis tel que la résolution des anaphores ([Mitkov 1998], [Mezghani 2008]), la vérification des erreurs d'accord ([Belguith 1999], [Makram Boujelben 2008]), etc.

3.4.3 Approches d'étiquetage grammatical

L'étiquetage grammatical est l'opération qui consiste à attribuer à chacun des mots d'un texte la catégorie (nom, verbe, adjectif, article défini, etc.) qui est la sienne dans le contexte où il apparaît [Debili 1998]. Ainsi, le but de l'étiquetage grammatical est de répondre à la question suivante : comment associer aux différents mots qui composent un texte l'étiquette qui leur convient, compte tenu du contexte où ils occurrent ?
Le principe de résolution le plus couramment utilisé repose sur une approche numérique qui fait intervenir des règles qui portent sur les successions permises ou non de deux, trois ou n étiquettes grammaticales. Toutefois, une formulation probabiliste est utilisée pour la résolution de l'ambiguïté dans le choix de la règle. Ainsi, les règles se sont vu adjoindre des poids statistiques afin de choisir les résolutions les plus probables. Ces règles peuvent être lues de plusieurs façons : on peut dire par exemple qu'après telle étiquette, ce sont telles ou telles étiquettes qui peuvent suivre ; mais si l'on considère la dernière étiquette, on peut également dire qu'elle dépend de celles qui la précèdent. Ainsi est-ce la formulation probabiliste utilisant les sources de Markov comme modèle qui s'est très vite répandue dès la fin des années 70 [Debili 2002].

3.4.4 Approches d'analyse morphologique

L'analyse morphologique a fait l'objet de plusieurs travaux de recherche qui sont classés généralement en deux approches :

- La première approche utilise un lexique comportant tous les mots (sous leurs différentes formes possibles) avec leurs caractéristiques associées. Dans ce cas, l'identification des caractéristiques morphologique des mots s'effectue par simple consultation du lexique. Cette approche suggère de stocker tous les mots avec leurs différentes formes. Dans ce cas, les dictionnaires se doivent de consigner toutes les unités attestées possibles, simples et construites. Ce type de lexique est appelé lexique en extension.
- La deuxième approche réduit le lexique aux seules informations non calculables (exemple : formes canoniques, racines, etc.) et utilise des règles mettent en jeu un savoir linguistique dans les traitements pour connaître le reste des informations. Cette approche consiste à choisir un lexique en intension, c'est-à-dire le noyau du lexique dénué de toute redondance pouvant servir de base à l'inventaire complet des lexèmes de la langue si on se donne un ensemble approprié de règles morphologiques.

Nous situons, dans ce qui suit, quelques systèmes d'analyse morphologique de l'arabe tels que le système ARAMORPH de Tim Buckwalter, l'analyseur Multi-Mode Morphological Processor (MMMP) de Sakhr, l'analyseur de XEROX de Kenneth Beesley et le système MORPH-2 de Chaâben qui a été réalisé au sein du laboratoire MIRACL. ARAMORPH, est un analyseur morphologique réalisé par Tim buckwalter [Buckwalter 2004] [17], qui utilise en entrée un texte qui doit être translittéré en ASCII avant tout traitement, et le résultat doit être re-converti en arabe pour qu'il soit compréhensible. Ce système ne permet pas l'analyse des textes contenant des chiffres et il se limite uniquement à l'analyse des mots qui figurent dans les dictionnaires arabes.

L'analyseur morphologique MORPH-2 de textes arabes non voyellé a comme objectif la détermination des caractéristiques morpho-syntaxiques de chaque mot composant le texte [18].

MORPH-2 utilise un lexique constitué de 3266 racines trilitères et quadrilitères de la langue arabe et repose sur cinq étapes à savoir : la segmentation de la phrase en mots, le prétraitement morphologique, l'analyse affixale, l'analyse morphologique et le post-traitement [Chaaben 2004].

La première étape est une étape de pré-traitement permettant de segmenter la phrase en mots.

La deuxième étape est aussi une étape de pré-traitement mais elle s'effectue au niveau du mot. Elle permet de supprimer les enclitiques et les proclitiques agglutinés au mot pour obtenir sa forme minimale. Ensuite, une étape de filtrage permettra l'élimination des particules de la liste des mots qui vont subir l'analyse affixale.

La troisième étape permet la détermination des préfixes, infixes et suffixes du mot, ainsi que la racine, s'il s'agit d'un verbe, ou la forme canonique s'il s'agit d'un nom.

La quatrième étape consiste à déterminer, à partir de chaque combinaison [19] , les caractéristiques morpho-syntaxiques correspondantes et de les affecter au mot correspondant.

La cinquième et dernière étape consiste à déterminer les groupements de mots possibles dans la phrase [Chaaben 2004].

Ce système a été évalué sur un corpus de textes non voyellés pris d'un livre scolaire tunisien (composé de 81 textes formant 899 paragraphes et 29 188 mots) et a donné un taux de rappel global de 69,77 % et un taux de précision global de 68,51% [Belguith 2008].

3.4.5 Approches de reconnaissance des entités nommées

La reconnaissance des entités nommées, s'intéresse à l'identification et le typage des noms de personnes, organisations, endroits, etc [Zribi 2010]. Ainsi, la reconnaissance des entités nom-

17. ARAMORPH : analyseur morphologique, téléchargeable à partir du site de LDC à l'adresse : http://www.nongnu.org/aramorph/french/

18. Au niveau de l'analyse morphologique, on ne s'intéresse pas au rôle du mot dans la phrase (sujet, complément, verbe)

19. i.e. préfixe, infixe, suffixe, racine/ forme canonique valide

mées est un problème qui se pose dans les différents domaines du traitement automatique des langues naturelles (TALN) à savoir : l'indexation de textes, la traduction de textes, l'extraction d'information, etc.
Dans le même ordre d'idées, Il est a signaler qu'il y a très peu de travaux visant la reconnaissance des entités nommées pour les textes arabes, et à notre connaissance il n'existe que les travaux réalisés par Benajiba [Benajiba 2009] et Mesfar [Mesfar 2008].
La plupart des travaux sur d'extraction d'entités nommées ont été créés généralement soit à base de réglés, soit à base d'apprentissage automatique.

- Les systèmes à base de règles sont généralement créés à partir de règles faites à la main. Ils se basent principalement sur la description des EN grâce à des règles qui exploitent un étiquetage syntaxique, des marqueurs lexicaux et des dictionnaires de noms propres.
  L'étiqueteur syntaxique reçoit en entrée un texte et produit automatiquement en sortie une version étiquetée de ce même texte. L'étiquetage consiste à produire les propriétés grammaticales d'un mot ou d'un groupe de mots dans une phrase donnée (e.g. : noms, verbes, conjonctions) souvent accompagnée d'informations morphologiques (e.g. : genre, nombre, personne).
  Les marqueurs lexicaux sont aussi appelés mots amorces ou mots déclencheurs ou parfois preuves contextuelles. Il s'agit de mots ou d'indices qui entourent (à gauche ou à droite) le nom propre et qui permettent souvent de prédire sa présence.
  Les règles utilisées dans ces systèmes sont décrites par des expressions de remplacement ou des expressions régulières (Regular Expression en anglais). Ainsi, la création des règles est basée sur des marqueurs lexicaux et elle dépend généralement de l'intuition du linguiste informaticien responsable de la création de ces règles. Plusieurs opérations et tests doivent être effectués afin de s'assurer de l'efficacité des règles.
  Parmi les systèmes à citer à base de règles : le système FUNES [Coates-Stephens 1992], le système Nominator [Wacholder 1997], etc.
- Les systèmes à base d'apprentissage sont conçus pour avoir une certaine "intelligence" lors de la prise des décisions. Ils se distinguent ainsi des systèmes à base de règles qui ne font qu'appliquer les règles injectées préalablement. Pour entraîner un système, il faut d'abord, lui fournir des données, appelées couramment corpus d'entraînement, et par la suite appliquer un algorithme d'apprentissage qui va permettre de construire automatiquement une base de connaissances. Le système sera alors prêt à fonctionner et à prendre des décisions d'une manière "autonome".
  Ainsi, Plusieurs approches d'apprentissage automatique ont été développées depuis quelques années. Nous pouvons citer par exemple le travail de Meulder et Daelemans [Daelemans 2003] ont employé le système d'apprentissage TIMBL afin de repérer les noms de personnes dans des journaux allemands et anglais. Ils ont combiné un corpus d'apprentis-

sage avec un lexique. Les premiers résultats ont montré que le lexique n'a apporté aucune amélioration des performances sur l'anglais ; par contre, il a pu améliorer les résultats sur l'allemand.

3.5 Conclusion

Dans ce chapitre, nous avons essayé de cerner les principaux problèmes d'analyse automatique de l'arabe inhérents à certains phénomènes tels que la non voyellation, l'agglutination, l'irrégularité de l'ordre des mots et l'absence de ponctuation régulière. Ensuite, nous avons donné un bref survol de l'état de l'art sur les principales approches et travaux réalisés dans le cadre de l'analyse automatique de l'arabe écrit.
Dans le chapitre suivant, nous allons présenter l'objet de notre étude, à savoir l'analyse rhétorique. Ainsi, nous commençons par proposer une étude analytique réalisée sur un corpus d'étude qui nous a permis de déduire, suite à des observations empiriques, un ensemble de relations, de frames (règles ou patrons) rhétoriques et morphologiques. Ensuite, nous décrirons dans un second temps l'analyse rhétorique qui a pour but d'établir les relations et les dépendances ainsi que l'importance relative à des phrases ou propositions les unes par rapport aux autres.

Chapitre 4

Annotation rhétorique

Sommaire

4.1 Introduction 63
4.2 Gestion des connaissances linguistiques 64
4.2.1 Construction du corpus 64
4.2.2 Étude du corpus 65
4.2.2.1 Relations rhétoriques 66
4.2.2.2 Frames rhétoriques basés sur des indices linguistiques 68
4.2.2.3 Frames rhétoriques basés sur des critères morphologiques 69
4.2.3 Organisation des frames rhétoriques en relations 71
4.2.4 Règles de correction des relations rhétoriques 72
4.3 Étapes de la méthode d'annotation rhétorique 73
4.3.1 Détermination de la relation rhétorique et de la nature de l'unité minimale 75
4.3.2 Enrichissement des relations rhétoriques 75
4.3.3 Correction des relations rhétoriques 75
4.3.4 Détermination de l'arbre RST le plus descriptif 76
4.4 Conclusion 78

4.1 Introduction

Le processus de résumé automatique de textes se base, comme nous l'avons déjà présenté précédemment dans le chapitre 2, sur une approche numérique et/ou symbolique. Il faut noter que ces deux approches, malgré leurs utilisations fréquentes dans les systèmes de résumé de textes indo-européens (l'anglais, le français, etc.) ont quelques limites. Cela est dû à la nature des corpus sur lesquels elles sont appliquées, au style et objectifs de l'auteur, etc.
Nous allons essayer dans ce qui suit de mentionner le potentiel d'efficacité apporté par l'annotation rhétorique dans l'amélioration de la qualité du résumé produit automatiquement afin de permettre à un utilisateur de répondre à ses besoins, notamment en l'aidant à se décider pour certaines actions.

La méthode d'annotation rhétorique, que nous proposons, se base sur la Théorie de la Structure Rhétorique (RST) [Mann 1988] et utilise des connaissances purement linguistiques. Ces connaissances sont principalement des frames rhétoriques, porteurs des valeurs sémantiques et indépendantes d'un domaine particulier. Les frames sont basés sur des indices linguistiques et des critères morphologiques, qui signalent des relations de discours (appelées relations rhétoriques) liant des unités minimales [1] contenues dans les textes.
La méthodologie suivie dans ce chapitre s'appuie sur deux piliers. Le premier pilier consistera à présenter le corpus de travail, l'analyser linguistiquement en se basant sur la validation de deux experts du domaine (linguistes spécialisés en langue arabe). Il s'agit d'une étape incontournable et qui se résume dans la recherche, la détection et l'organisation des frames rhétoriques et des règles morphologiques trouvées. Le second pilier sera consacré à présenter notre proposition afin de détecter, d'enrichir et de corriger l'étiquetage rhétorique qui conduit à la fin à la construction de l'arbre RST du texte [Maaloul 2010a].

4.2 Gestion des connaissances linguistiques

Les études consacrées à la RST ont intéressé plusieurs chercheurs comme Ballard et al. [Ballard 1971], Halliday et Hasan [Halliday 1976], Longacre [Longacre 1979], Hovy [Hovy 1998], Mathkour et al. [Mathkour 2008], etc., qui considèrent que le choix du corpus de travail et son analyse linguistique sont un facteur privilégié pour déterminer les critères de détection des relations sémantiques et des relations intentionnelles qui existent entre les segments d'un discours. Pour établir cette cohérence, une analyse linguistique des discours formant notre corpus d'étude s'avère nécessaire et primordiale.
Ainsi, le but essentiel de cette analyse est premièrement i) le repérage des unités linguistiques de surface qui représentent des marqueurs linguistiques indépendants d'un domaine particulier et porteurs de valeurs sémantiques communes aux articles de presse, et deuxièmement ii) l'organisation de ces marqueurs dans des relations rhétoriques.
Afin d'illustrer les particularités de notre domaine d'étude, nous allons présenter les différentes étapes de notre étude sur des articles de presse.

4.2.1 Construction du corpus

L'utilisation des corpus dans les applications du TALN n'est pas récente. Dès la fin du XX^{me} siècle, l'analyse manuelle de corpus a permis l'étude de la fréquence d'occurrence des mots et des graphèmes ainsi que l'extraction des listes de mots et des collections dans le but

1. Les auteurs de la RST définissent les unités minimales comme des unités fonctionnellement indépendantes : elles correspondent généralement aux propositions [Mann 1988].

d'étudier la faculté d'acquisition du langage en enseignement [Ide 1996].
Au début du XXI^{me} siècle, les corpus jouaient un rôle important dans le domaine de la linguistique computationnelle. Ce rôle a vu un affaiblissement pendant les milieux des années cinquante à cause de la critique chomskyenne [2] dans [Manguin 2003]. Par conséquent, peu sont les linguistes qui ont continué leurs travaux à base de corpus.
Dans les années 1980, l'utilisation de corpus en TALN connaît une vive renaissance (pour plusieurs langues naturelles et spécialement pour l'anglais), et l'on voit même apparaître une linguistique dite le "*corpus*". Le succès des méthodes empiriques dans la reconnaissance de la parole a entraîné un nouvel accroissement de l'intérêt des corpus, pratiquement dans toutes les applications du traitement automatique des langues.
Afin de mieux se situer dans ce cadre, nous avons choisi d'utiliser un corpus d'étude auquel nous ferons référence dans la suite de ce rapport. Ce dernier est formé par des articles de presse de type HTML avec un codage UTF-8 (Unicode Transformation format). Le choix de ce type de document se justifie par son accessibilité sur le Net.
En effet, c'est à partir du Web que nous avons construit notre corpus de textes en langue arabe [3]. Ces articles de presse ont été rapatriés sans restriction quant à leur thème (Éducation, Sport, Science, Politique, Reportage, etc.), leur contenu et leur volume. La raison c'est que nous estimons que plus le corpus est varié, plus il sera représentatif ; par conséquent le nombre de marqueurs linguistiques qu'il contiendra sera plus important [Maaloul 2010d].
La taille moyenne des articles formant le corpus est de quatre pages. Quatre milles articles sont utilisés pour la réalisation et cent articles sont utilisés pour l'évaluation.
Ce corpus servira, plus tard, comme source d'extraction de marqueurs linguistiques utiles pour la tâche de l'analyse sémantique d'un article de presse.

4.2.2 Étude du corpus

Le but de cette section vise à appliquer une analyse linguistique de surface à notre corpus. Cette analyse consiste à mettre en évidence des connaissances purement linguistiques formées principalement par des marqueurs linguistiques et des indicateurs (*features*) qui ont des valeurs importantes dans un article de presse et qui assistent le traitement linguistique en couvrant différents cas d'interaction entre les segments du texte.
Ainsi, Nous avons entrevu tout au long de cette étape d'étude du corpus deux experts qui nous

2. En 1959, Chomsky publie une critique marquante du livre de Burrhus Skinner Verbal Behavoir cité dans [Véronis 2000]. Dans son livre, Skinner explique son point de vue comportementaliste du langage. Le comportement linguistique y est défini comme un comportement appris, avec pour conséquence d'être transmis par le comportement déjà appris par d'autres individus citée

3. Source : site électronique `http://www.daralhayat.com` est un journal quotidien généraliste arabophone avec une diffusion dans le monde arabe de 110 000 exemplaires.

ont déterminé des éléments linguistiques formés principalement de marqueurs linguistiques indépendants d'un domaine particulier. Ceux-ci sont les traces directes de l'intention énonciative de l'auteur du texte et les instruments qu'il utilise pour guider le lecteur dans son processus cognitif de compréhension [Alrahabi 2008].
Nous commençons notre étude par l'analyse sémantique qui consiste à repérer les traits, étiquettes sémantiques (*relations rhétoriques*), caractérisant le contenu essentiel d'un article. Ces étiquettes sémantiques permettent de donner des valeurs sémantiques de l'information à retenir dans un texte. Ainsi, notre analyse sémantique du corpus est réalisée par une lecture sélective afin de filtrer les segments textuels correspondants à ces étiquettes.
Toutefois, ces marqueurs peuvent être répertoriés en deux types : *indicateurs déclencheurs* et *indices complémentaires* [Minel 2002a]. En effet, l'objectif de cette étude fût de déceler, dans un premier temps, des indicateurs déclencheurs énonçant des concepts importants et qui sont pertinents pour la tâche de résumé automatique ; et ensuite à rechercher des indices complémentaires dans un espace de recherche défini à partir de cet indicateur (dans le voisinage de l'indicateur) [Alrahabi 2006a].
Ces *indicateurs déclencheurs* et *indices complémentaires* forment les *frames* (*règles* ou *patrons*) de *relations rhétoriques*. Les frames sont des règles rhétoriques formées par des signaux linguistiques et des heuristiques observées qui sont principalement des marqueurs indépendants d'un domaine particulier, mais qui ont des valeurs importantes dans un article de presse [Alrahabi 2009].
Ainsi, les marqueurs (indicateurs déclencheurs et indices complémentaires) formant les frames rhétoriques ont un double rôle : premièrement ils constituent un moyen de témoignage nécessaire pour lier deux *unités minimales* adjacentes, dont l'une possède le statut de noyau – segment de texte primordial pour la cohérence – et l'autre a le statut *noyau* ou *satellite* – segment optionnel [Luc 2001] ; deuxièmement ils permettent la construction des différents arbres rhétoriques possibles [Maaloul 2010a].
À l'aide de ces indicateurs et des informations qu'ils portent, on peut découper le texte en différentes unités textuelles à partir desquelles nous allons extraire, par le moyen d'autres critères, les unités les plus pertinentes. Ces unités textuelles, appelées aussi segments, vont être pondérées selon leur appartenance à une relation rhétorique donnée.

4.2.2.1 Relations rhétoriques

Comprendre un texte écrit, c'est mobiliser à la fois une compétence linguistique et un savoir interprétatif à partir des indices fournis par le texte lui-même [Pascual 1995].
Dans ce cadre d'hypothèse, notre interprétation analytique par l'analyse sémantique menée sur des textes du corpus nous a permis de dégager des relations rhétoriques qui ont pour but de rattacher des segments distincts et contigus du texte source.

Signalons que dans la littérature, et en ce qui concerne les relations rhétoriques de discours, il n'existe pas de liste définitive des relations adoptées pour les différents types de discours. Pour l'analyse de discours narratifs, Asher et Lascarides [Asher 2003] proposent les relations suivantes :

TABLEAU 4.1 – Liste des relations rhétoriques définies par Asher et Lascarides pour un discours narratif [Asher 2003]

"Narration"	"Continuation"	"Résultat"	"Arrière-plan"
"Parallèle"	"Élaboration"	"Acquiescement	"Topique"
"Conséquence"	"Explication"	"Contraste"	"Plan-Élab"

À l'issue de notre étude empirique menée sur un corpus d'étude, nous avons pu obtenir vingt relations rhétoriques validées par deux experts. La validation des relations rhétoriques est faite suite à un accord total entre les deux experts sur deux mille cinq cent quatre-vingts exemples de phrases porteuses des relations rhétoriques différentes.
Ces relations rhétoriques nous ont permis de décrire une vision accointance des unités fonctionnellement indépendantes (généralement ce sont des propositions) du dialogue et de renfermer ces différentes unités du texte source sous un format organisationnel d'une façon hiérarchique.
Il est à signaler que suite à des contraintes méthodologiques, et afin de pouvoir construire l'arbre RST à la fin de l'analyse rhétorique, nous avons ajouté une relation rhétorique "Autres - آخر". Cette relation est attribuée lorsqu'aucune relation rhétorique n'est déterminée dans une phrase, nous parlons-ici d'une phrase non porteuse d'information rhétorique.
À la fin de cette étude, nous avons pu énumérer les relations rhétoriques présentées dans le tableau suivant 4.2 :

TABLEAU 4.2 – Liste des relations rhétoriques

"Autres - آخر"	"Exception - استثناء"	"Concession - استدراك"	"Explication - تفسير"
"Condition - شرط"	"Pondération - ترجيح"	"Énumération - تفصيل"	"Classement - ترتيب"
"Réduction - تقليل"	"Restriction - حصر"	"Exemplification - تمثيل"	"Évidence - قاعدة"
"Joint - ربط"	"Définition - تعريف"	"Confirmation - توكيد"	"Négation - نفي"
"Possibilité - إمكان"	"Justification - تعليل"	"Spécification -تخصيص"	"Conclusion - استنتاج"
"Affirmation - جزم"			

Notons que quelques unes de ces relations rhétoriques sont communes avec celles trouvées par Waleed [Mathkour 2008].

4.2.2.2 Frames rhétoriques basés sur des indices linguistiques

Les frames sont des règles rhétoriques formées par des signaux linguistiques et des heuristiques observés qui sont principalement des marqueurs indépendants d'un domaine particulier mais qui ont des valeurs importantes dans un article de presse [Alrahabi 2006a].
Rappelons que ces frames rhétoriques ont pour objectifs la distinction entre les unités minimales noyaux et satellites, outre la détermination des relations rhétoriques. En effet, une unité minimale noyau exprime ce qui est plus essentiel au but de l'auteur que le satellite ; et que le noyau d'une relation rhétorique est indépendant compréhensible du satellite, mais pas vice-versa.
Toutefois, la détection d'une relation se fait généralement par l'application d'un ensemble de contraintes sur le noyau, le satellite et sur la combinaison du noyau et du satellite.
L'exemple suivant illustre une phrase repérée dans l'un des articles de notre corpus :

(أ)لكن ألبير قصيري لم يكن نزيل غرفته في ذلك الفندق فقط، بل كَان أحد وجوه الشَارع وبعض مقَاهيهَا الشهيرة، (ب) لَا سيمَا مقهَى آفلوز الذي كَان يقضي فيه سَاعَات وحيدَاً أو مع أشخَاص عَابرين.

(O)lkn Olbyr qSyry lm ykn nzyl grfth fy *lk Alfndq fqT, bl kAn OHd wjwh Al$ArE wbED mqAhyhA Al$hyrp, (b) lA symA mqhY "flwr" Al*y kAn yqDy fyh sAEAt wHydAF Ow mE O$xAS EAbryn.
(A)Mais Albert Kasiry **n**'était **pas** seulement résident de sa chambre dans cet hôtel, **mais** il était l'un des gens connus dans la rue et de certains de ses cafés célèbres, (B) **notamment** le café "Flor" où il passe quelques heures tout seul ou avec des personnes passagères.
Cette phrase contient une relation de "Spécification - تخصيص" entre la première unité minimale (A) - (أ) et la deuxième unité minimale (B) - (ب). Une relation de "Spécification - تخصيص" a généralement pour rôle de détailler ce qui est indiqué et de confirmer son sens et de le clarifier.
Le tableau suivant 4.3 présente le frame utilisé pour détecter la relation rhétorique "Spécification - تخصيص" :

TABLEAU 4.3 – Exemple de frame utilisé pour la détection de la relation rhétorique "Spécification - تخصيص"

Relation :	"Spécification - تخصيص"
Contrainte sur (A) - (أ) :	contient un/des indice(s) complémentaire(s) "**mais** - بل", "**ne pas** - لم" , "**non** - لَا", etc.
Contrainte sur (B) - (ب) :	contient l'indice déclencheur "**ainsi** - لَاسيمَا".
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Le marqueur "**ainsi** - لَاسيمَا" est un indicateur déclencheur de recherche.
Les marqueurs "**mais** - بل" et "**ne pas** - لم" sont des indices complémentaires qui valident les conditions d'appartenance de l'indicateur "**ainsi** - لَاسيمَا" à sa relation rhétorique, "Spécification - تخصيص".

4.2.2.3 Frames rhétoriques basés sur des critères morphologiques

Les frames sont des faits rhétoriques qui ont comme objectif la reconnaissance d'une relation rhétorique tout en se reposant sur une interprétation morpho-syntaxique du contenu du segment.
Cette interprétation s'effectue suivant une analyse morphologique [4] du texte. L'interprétation résultante est donc un ensemble de relations rhétoriques permettant de lier deux segments adjacents entre eux, dont l'un possède le statut de noyau et l'autre celui de satellite.
Ces segments peuvent être de deux types : segment minimal (*Text Unit*) ou segment composé (*Text Span*). Toutefois, ce qui nous intéresse dans le cadre de notre étude est l'ensemble des formes basé sur des critères morphologiques utiles dans le repérage de relations rhétoriques.
Cependant, les observations réalisées sur le corpus d'étude ont montré un certain nombre d'ambiguïtés au niveau de la détermination de la bonne relation rhétorique reliant deux segments adjacents [Maaloul 2010d].
Les exemples 1 et 2 montrent que l'indicateur déclencheur "قد" énonce deux relations rhétoriques différentes : "Pondération - ترجيح" et "Affirmation - جزم". En absence d'indices complémentaires au voisinage de l'indicateur déclencheur la confirmation du concept énoncé par l'indicateur déclencheur est impossible.
Ainsi, pour lever cette ambigüité, nous proposons d'utiliser les caractéristiques morphologiques des mots du segment textuel (i.e., le type de la catégorie grammaticale des mots : nom, verbe, pronom, etc. et le temps pour les verbes et).
Les frames rhétoriques relatifs à ces deux exemples sont les suivantes :

Exemple 1 :

(أ) وفي هذَا الحَال الذي لَا يسر، كَانت ثمة تحذيرَات من قبل جنرَالَات سَابقين أن المصير الأَمريكي قد (ب) يذهب إِلَى مآلَات خطيرة..إِلخ.

(O) wfy h*A AlHAl Al*y lA ysr, kAnt vmp tH*yrAt mn qbl jnrAlAt sAbqyn On AlmSyr AlOmryky qd (b) y*hb IlY m|lAt xTyrp...Ilx.
(A) Dans ces conditions regrettables, il y a des avertissements de la part du général d'armée

4. L'analyse morphologique produit un ensemble d'informations (base, racine, catégorie grammaticale, ensemble de traits syntaxiques, etc) des mots du texte

que le destin de la politique américaine **peut** (B) conduire à des conséquences dangereuses . . . etc.

Exemple 2 :

(أ) مسؤول رفيع من النادي الملكي قد(ب) ذهب ألى فرنسا و راقب نجم ليون كريم بن زيمـا...

(O) msWwl rfyE mn AlnAdy Almlky qd (b) *hb AlY frnsA w rAqb njm lywn krym bn zymA ...Ilx.
(A) Un responsable de l'équipe royale **est** (B) allé en France pour observer la vedette lyonnaise "Karim Benzema"... etc.

TABLEAU 4.4 – Exemple de frame utilisé dans l'exemple 1, pour la détection de la relation rhétorique "Pondération - ترجيح"

Relation :	"Pondération - ترجيح"
Contrainte sur (A) - (أ) :	contient un/des indice(s) déclencheur(s) "**Il se peut que** - قد", "**Il a** - وقد" suivit d'un **un verbe au futur**
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

TABLEAU 4.5 – Exemple de frame utilisé dans l'exemple 2, pour la détection de la relation rhétorique "Affirmation - جزم"

Relation :	"Affirmation - جزم"
Contrainte sur (A) - (أ) :	contient un/des indice(s) déclencheur(s) "**Il a** - لقد", "**Il se peut que** - قد", "**Il a** - وقد" suivit d'un **un verbe au passé**
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

4.2.3 Organisation des frames rhétoriques en relations

Il s'agit dans cette phase d'ordonner les *frames rhétoriques basés sur des indices linguistiques* (formées par les marqueurs c'est-à-dire indicateurs déclencheurs et indices complémentaires) et *les frames rhétoriques basés sur des critères morphologiques* (formées par des étiquettes morphologiques c'est-à-dire catégories grammaticales et caractéristiques morphologiques), selon les relations rhétoriques.

Ainsi, nous aurons dans une relation rhétorique une liste de patrons linguistiques formés d'un ensemble d'unités linguistiques dont les catégories sont parfois hétérogènes (noms, verbes, connecteurs, mots outils ou grammaticaux, etc.) mais qui remplissent toujours les mêmes fonctions sémantiques discursives [Maaloul 2010b].

Les tableaux 4.6 et 4.7, présentent quelques exemples de frames basés sur des indices linguistiques réparties selon les relations rhétoriques.

Chaque frame est formé par le nom de la relation qu'il indique, son indicateur déclencheur, sa liste d'indices complémentaires de validation qui peuvent être rencontrés au voisinage de l'indicateur, la position du marqueur déclencheur indique qui peut être au début, au milieu ou à la fin de l'unité de texte et enfin l'unité(s) minimale(s) retenu(es) (le noyau). Le tableau 4.8,

TABLEAU 4.6 – Exemple de frame utilisé pour la détection de la relation rhétorique "Négation - نفي"

Relation :	"Négation - نفي"
Contrainte sur (A) - (أ) :	contient un/des indice(s) complémentaire(s) "**mais** - بل", "**d'ailleurs** - أمَا", "**cependant** - لكن", etc.
Contrainte sur (B) - (ب) :	contient un/des indice(s) déclencheur(s) "**revenir sur une décision** - لن", "**ne pas** - لم", "**pas** - ليس"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

présente un exemple de frame basé sur des critères morphologiques réparties selon les relations rhétoriques.

Ce frame est formé par le nom de la relation qu'il indique, son indicateur déclencheur, les caractéristiques morphologiques des mots des segments textuels (i.e. : le type de la catégorie grammaticale des mots et le temps pour les verbes) qui peuvent être rencontrés après l'indicateur déclencheur, la position du marqueur déclencheur indique qui peut être au début, au milieu ou à la fin de l'unité de texte et enfin l'unité(s) minimale(s) retenu(es) (le noyau).

TABLEAU 4.7 – Exemple de frame utilisé pour la détection de la relation rhétorique "Confirmation - توكيد"

Relation :	"Confirmation - توكيد"
Contrainte sur (A) - (أ) :	contient l'indice complémentaire "**prenant** - وَاذ"
Contrainte sur (B) - (ب) :	contient un/des indice(s) déclencheur(s) "**J'ai** - لقد", "**si seulement** - إن", "**bien que** - رغم"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

TABLEAU 4.8 – Exemple de frame utilisé pour la détection de la relation rhétorique "Pondération - ترجيح"

Relation :	"Pondération - ترجيح"
Contrainte sur (A) - (أ) :	contient un/des indice(s) déclencheur(s) "**Il se peut que** - قد", "**Il a** - وقد"**suivit** d'un **un verbe au futur**
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

4.2.4 Règles de correction des relations rhétoriques

À l'issue de notre processus d'annotation manuelle, nous avons remarqué que les problèmes relatifs à la richesse flexionnelle de la langue arabe et la voyellation résident dans l'absence quasi systématique de la voyellation dans les articles de presse de notre corpus d'étude ; ce qui provoque dans la plupart du temps des ambiguïtés au niveau de la détection de la bonne relation rhétorique qui relie deux unités minimales successives.
Notre analyse du corpus de travail, nous a permis de dégager diverses compositions des règles de correction des relations rhétoriques. Ces règles de correction peuvent être appliquées sur deux relations rhétoriques adjacentes ou entre un indice et une relation rhétorique.
L'algorithme suivant 6 présente un exemple de cas de correction entre deux relations rhétoriques. Ainsi, le but est de faire une recherche sur les relations rhétoriques adjacentes déjà déduites à l'aide des frames rhétoriques (basés sur des indices linguistiques et sur des critères morphologiques) afin de détecter une relation de "Confirmation - توكيد" suivie d'une relation de "Négation - نفي" et de remplacer ces deux relations adjacentes par une nouvelle relation

"d'Affirmation - جزم".

Remarquons que l'indicateur déclencheur de la relation "Négation - نفي" sera dans ce cas le délimiteur de la nouvelle relation "d'Affirmation - جزم", où l'unité minimale qui le précède et qui le suit est considérée comme une unité Noyau.

Algorithme 6 Exemple d'une règle de correction de type relation-relation

1: **soit** le contexte C : une phrase
2: **if** (L'on repère la relation "Confirmation - توكيد" suivie de la relation "Négation - نفي" dans C) **then**
3: Les deux relations seront remplacées par une relation "d'Affirmation - جزم" **et**
4: L'unité minimale qui précède l'indicateur déclencheur de la relation "Négation - نفي" est considérée comme une unité minimale Noyau **et**
5: L'unité minimale qui suit l'indicateur déclencheur de la relation "Négation - نفي" est considérée comme une unité minimale Noyau

Dans l'exemple suivant 7 nous allons voir le deuxième type de règle de correction qui sera appliqué suite à une recherche d'un indice suivi d'une relation.

Algorithme 7 Exemple d'une règle de correction de type indice-relation

1: **soit** le contexte C : une phrase
2: **if** (L'on repère la relation "Négation - نفي" suivie de l'indice "Parce-que - لأَن" dans C) **then**
3: La relation "Négation - نفي" sera remplacée par une relation "Explication - تفسير" **et**
4: L'unité minimale qui précède l'indice "Parce-que - لأَن" est considérée comme une unité minimale Noyau **et**
5: L'unité minimale qui suit l'indice "Parce-que - لأَن" est considérée comme une unité minimale Satellite

En consultant les deux derniers tableaux, nous remarquons qu'il existe deux types de combinaisons possibles pour les règles de correction.

4.3 Étapes de la méthode d'annotation rhétorique

La méthode que nous proposons pour la modélisation du texte source sous une forme rhétorique enchâssée, est basée sur la théorie des structures rhétoriques [Mann 1988]. Ainsi, notre méthode d'annotation rhétorique se base sur une technique purement symbolique qui a comme but d'établir les relations et les dépendances ainsi que l'importance relative des phrases ou

propositions les unes par rapport aux autres [Teufel 1998].

En effet, l'analyse rhétorique vise à donner une illustration linguistique cohérente aux différentes parties des textes, afin de détecter les relations sémantiques et les relations intentionnelles qui existent entre les segments d'un texte.

Le but de cette analyse consiste à :

- la détection des relations rhétoriques qui existent entre les différentes unités minimales adjacentes d'une phrase dont l'une possède le statut de noyau – segment de texte primordial pour la cohérence – et l'autre a le statut noyau ou satellite – segment optionnel.
- l'enrichissement des relations rhétoriques inférées à partir des frames rhétoriques basés sur des critères linguistiques (voir 4.2.2.2) par d'autres relations déduites à partir des frames rhétoriques basés sur critères morphologiques (voir 4.2.2.3).
- la correction des différentes relations détectées après la première et la deuxième phases. Cette phase a pour objet de réduire les cas d'ambiguïtés dans le choix des relations rhétoriques d'attachement entre les différentes unités minimales (on privilégie l'augmentation de la cohérence en appliquant un ensemble de règles de correction de type relation-relation et marqueur-relation (voir 4.2.4)).
- la détermination de l'arbre RST le plus descriptif du texte, afin de spécifier la composition structurale du texte, en se basant sur des schémas rhétoriques et sur un certain nombre de réglés rhétoriques [Maaloul 2011].

La figure 4.1 présente les principales phases de l'analyse rhétorique.

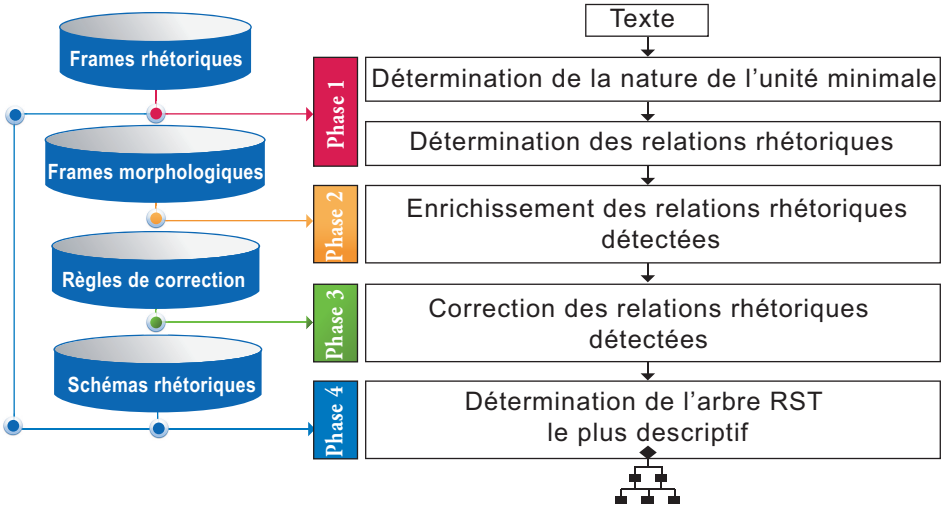

FIGURE 4.1 – Phases d'analyse rhétorique

Dans ce qui suit, nous décrivons les différentes phases composant l'analyse rhétorique du texte source.

4.3.1 Détermination de la relation rhétorique et de la nature de l'unité minimale

Cette phase a un double objectif ; premièrement de lier deux unités minimales adjacentes entre elles, dont l'une possède le statut de noyau et l'autre a le statut noyau ou satellite, et deuxièmement de déterminer les relations rhétoriques qui existent entre les différentes unités minimales juxtaposées d'un même paragraphe [Maaloul 2011].
Il est à signaler que cette phase de détermination des relations rhétoriques et de la nature des unités minimales se base sur une analyse de surface (technique d'exploration contextuelle - voir 2.3.5) et ce en utilisant seulement les frames rhétoriques basés sur des critères linguistiques (indicateurs déclencheurs de recherche et indices de validation).

4.3.2 Enrichissement des relations rhétoriques

La deuxième phase d'enrichissement des relations rhétoriques consiste à appliquer des frames rhétoriques basés sur des critères morphologiques afin de déduire d'autres relations différentes de celles déjà repérées lors de la première phase.
Ces règles morphologiques sont des règles rhétoriques formées d'un indicateur déclencheur de la relation rhétorique et d'un ensemble de critères morphologiques des mots au voisinage de l'indice déclencheur (les catégories de ces mots sont parfois hétérogènes : noms, verbes, connecteurs, mots outils ou grammaticaux, etc.).
Un étiquetage morphologique du texte source s'avère nécessaire et primordial avant d'entamer cette phase d'enrichissement, vu que l'application des règles d'enrichissement (qui sont des frames rhétoriques basés sur des critères morphologiques - voir 4.2.2.3) nécessitent ce type d'informations morphologiques.

4.3.3 Correction des relations rhétoriques

Cette phase de correction consiste principalement à appliquer les règles de correction sur l'ensemble des relations rhétoriques déterminées par la première et la deuxième phase. Rappelons, que ces règles de correction sont présentées sous la forme de deux types de règles : indice-relation et relation-relation.
L'objectif principal de cette phase de correction outre la correction des relations rhétoriques, c'est que nous pourrons nous servir de cette information de correction (les nouvelles relations rhétoriques) à la prochaine phase afin d'offrir un seul arbre RST.
Ainsi, l'idée de construire plusieurs arbres RST et de choisir celui le plus descriptif est éliminé vu qu'il existe, après la phase de correction, seulement une seule relation qui relie deux unités minimales juxtaposées.

4.3.4 Détermination de l'arbre RST le plus descriptif

Cette phase consiste à créer l'arbre RST le plus descriptif qui décrit l'organisation structurelle du texte source, et cela en prenant en considération les différentes contraintes des liens entre unités minimales du texte. Ainsi, l'arbre RST se caractérise par sa capacité de connecter récursivement, par le biais d'une relation de discours qui ne peut relier que des segments de texte adjacents, les unités minimales et les segments de texte plus larges ainsi construits selon un ordre cohérent et informatif.
Il est à signaler que cette phase de création et de détermination de l'arbre RST le plus descriptif a fait l'objet de plusieurs travaux de recherche. Parmi ces travaux nous pouvons citer les recherches de Marcu. Selon Marcu et suite à ces études empiriques, l'arbre RST le descriptif est celui le plus équilibré à droite et à gauche [Marcu 2000b].
Toutefois, l'approche de la sélection de l'arbre le plus descriptif au texte reste au niveau théorique car une telle représentation ne peut pas être obtenue généralement de manière automatique [Marcu 2000b].
Ainsi, nous proposons l'utilisation d'une technique qui fait appel à un certain nombre de règles et de schémas rhétoriques afin de décrire l'organisation structurelle d'un texte, quel que soit son niveau hiérarchique.
Les règles rhétoriques sont utilisées afin d'hiérarchiser et d'affiner l'arbre RST. Elles utilisent des heuristiques, adoptées après observation des résultats. Nous donnons ici à titre représentatif une règle rhétorique [Maaloul 2010a].

Algorithme 8 Exemple de règle rhétorique utilisée pour la construction de l'arbre RST

1: **if** (Un **indicateur déclencheur** se trouve au **début de phrase**) **then**
2: La phrase annotée est en relation avec le passage qui la précède.

Pour les schémas rhétoriques, ils se présentent sous la forme de cinq modèles de schémas et qui peuvent être utilisés récursivement quel que soit le niveau hiérarchique de ce dernier. Ils permettent , ainsi, de lier un noyau et un satellite, deux ou plusieurs noyaux entre eux, et un noyau avec plusieurs satellites [Marcu 2000b] afin décrire une structure rhétorique finale d'un texte, de taille arbitraire, et strictement hiérarchique et se présente sous la forme d'un arbre RST.
L'exemple suivant présente une interprétation RST (voir figure 4.2) déduite à partir des modèles de schémas présentés précédemment relatifs au paragraphe suivant.

(أ) تشتهر مدينة صفَاقس بتقديم أطبَاق ثمَار البحر علَى أنوَاعهَا. (ب) عندمَا يرتَاد زوَار مدينة صفَاقس، فَانهم يطلبون بَاستمرَار أطبَاق ثمَار البحر (ت) وخَاصة طبق المحَار وَالأَخطبوط المشوي علَى الفحم.

(O) t$thr mdynp SfAqs btqdym OTbAq vmAr AlbHr ElY OnwAEhA. (b) EndmA yrtAd zwAr mdynp SfAqs, fInhm yTlbwn bAstmrAr OTbAq vmAr AlbHr (t) wxASp Tbq AlmHAr wAlIxTbwT Alm$wy ElY AlfHm.
(A) La ville de Sfax est connue par la présentation des plats de fruits de mer de tout type. (B) **Lorsque** les visiteurs se rendent à la ville de Sfax, **ils** demandent régulièrement les plats de fruits de mer (C)**et surtout** le plat d'huître et de poulpe grillé sur le charbon.

Il est à noter que le jugement d'appartenance à la relation rhétorique "Évidence - قَاعدة" est attribué aux unités minimales "(A) - (أ)" et "(B) - (ب)". Cette attribution est faite en se basant sur le frame rhétorique suivant :

TABLEAU 4.9 – Exemple de frame utilisé pour la détection de la relation rhétorique "Évidence - قَاعدة"

Relation :	"Évidence - قَاعدة "
Contrainte sur "(A) - (أ)" :	-
Contrainte sur "(B) - (ب)" :	contient l'indice déclencheur "**Lorsque** - عندمَا".
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Alors que la relation rhétorique "Condition-شرط" est attribuée aux unités minimales "(B) - (ب)" et "(C)-ت".
Cette attribution est faite en se basant sur le frame rhétorique composé d'un indicateur déclencheur de recherche "**et surtout**-وخَاصة" et l'indice complémentaire "**ils**-فَانهم".

TABLEAU 4.10 – Exemple de frame utilisé pour la détection de la relation rhétorique "Condition - شرط"

Relation :	"Condition - شرط"
Contrainte sur "(B) - (ب)" :	contient l'indice complémentaire "**ils** - فَانهم".
Contrainte sur "(C) - ت" :	contient l'indice déclencheur "**et surtout** - وخَاصة".
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

La RST va réagir à cet exemple comme suit en appliquant la règle 8 et nous aurons comme résultat l'arbre suivant :

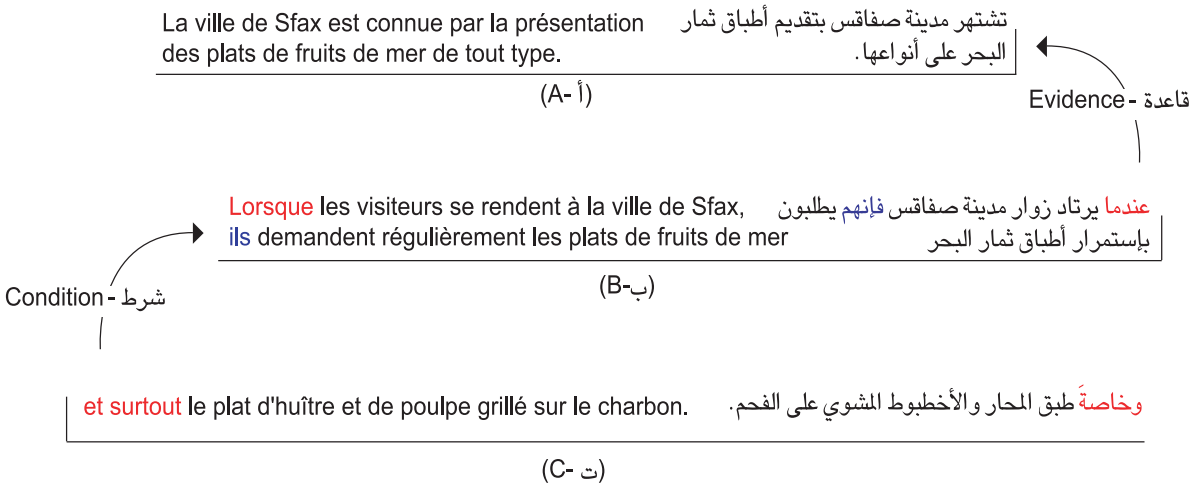

FIGURE 4.2 – Arbre RST

Dans le reste de ce travail, nous nous appuyons principalement sur la distinction Noyau/Satellite prônée par la RST et qui, selon nous, instaure une relation de dépendance entre deux unités minimales.

4.4 Conclusion

Dans le présent chapitre, nous avons explicité les améliorations que nous avons proposées au niveau de la théorie de la structure rhétorique (RST) classique définie par [Mann 1988]. Nous avons, ainsi, intégré deux phases d'enrichissement et de correction afin de déterminer un seul arbre descriptif.

Au niveau de l'enrichissement des relations rhétoriques, nous avons proposé l'utilisation des frames rhétoriques. Ces frames se basent sur des indicateurs déclencheurs de relations et un ensemble de critères morphologiques permettent de lever l'ambiguïté dans certains cas où on trouve un indicateur déclencheur de recherche sans indices de validation.

Au niveau de la correction, notre contribution s'est manifestée par la proposition d'un ensemble de règles de correction en vue de réduire le nombre de relations rhétoriques. Cette correction se base sur des règles de type relation-relation et indice-relation.

Pour générer l'arbre final d'un texte, nous nous sommes basés sur les cinq schémas définis par Mann et Thompson [Mann 1988] et vingt-cinq frames rhétoriques [Maaloul 2010a].

Dans le chapitre suivant, nous allons, détailler les étapes de notre proposition qui consiste à générer un extrait par une approche hybride, à savoir : la segmentation, l'étiquetage morphologique, l'analyse rhétorique et le classement et sélection des phrases selon le type du résumé.

Chapitre 5

Génération d'extrait par une approche hybride

Sommaire

5.1 Introduction 79
5.2 Notre proposition : étapes de l'approche proposée 80
5.2.1 Segmentation du document source 81
5.2.2 Étiquetage morphologique 82
5.2.3 Analyse rhétorique 83
5.2.4 Phase d'apprentissage 84
5.2.4.1 Corpus d'apprentissage 85
5.2.4.2 Apprentissage basé sur l'algorithme SVM 87
5.2.4.3 Vecteur d'extraction SVM 88
5.2.4.4 L'algorithme SMO 89
5.2.4.5 Scénario d'application de l'algorithme SVM 90
5.2.5 Sélection et classement des phrases selon le type du résumé 90
5.2.5.1 Classement et sélection des phrases 90
5.2.5.2 Vers un résumé dynamique 91
5.3 Conclusion 92

5.1 Introduction

Dans le chapitre précédent, nous avons présenté l'élément de base dans notre processus de génération d'extrait qui est l'annotation rhétorique. Cette annotation se distingue par sa capacité de produire un seul arbre RST qui connecte récursivement les unités minimales et les segments de texte plus larges ainsi construits selon un ordre cohérent et informatif.
Selon cette vision, notre proposition pour la génération d'extrait opère par l'utilisation d'une approche hybride qui combine une analyse purement symbolique avec une sélection purement numérique. Pour illustrer cet aspect hybride, nous proposons d'utiliser une analyse rhétorique

pour déterminer les relations rhétoriques et une technique d'apprentissage afin de sélectionner des unités textuelles (phrases) formant l'extrait final. Cependant, la technique d'apprentissage tient en compte les relations rhétoriques mentionnées par l'arbre RST, le type de résumé choisi et les besoins potentiels d'un utilisateur.
dans à cet ordre d'idée, nous mentionnons dans ce chapitre qu'une information n'est pas importante en soi, mais doit correspondre aux besoins d'un utilisateur. Ainsi, l'approche hybride, que nous proposons, aborde la question des besoins des utilisateurs. En effet, notre recherche s'oriente de plus en plus vers la production de résumés dynamiques [Maâloul 2008].
Nous allons aussi détailler les étapes de l'approche hybride proposée et pour mener au mieux cet aspect d'hybridation nous commençons par décrire globalement notre proposition, puis nous détaillons chacune de ses étapes : la segmentation, l'étiquetage morphologique, l'analyse rhétorique, l'apprentissage, le classement et la sélection des phrases selon le type de résumé.

5.2 Notre proposition : étapes de l'approche proposée

Dans notre proposition, l'extraction automatique des passages pertinents du texte est générée par instanciation d'une approche basée sur l'utilisation conjointe d'un traitement numérique et d'une analyse symbolique.
L'analyse symbolique consiste en une analyse rhétorique (voir chapitre 4) qui a pour rôle de mettre en évidence dans un texte source les unités minimales (noyaux – segments de texte primordiaux pour la cohérence et qui charpentent un discours) nécessaires pour le processus d'extraction.
Cette analyse symbolique fait appel, en seconde phase, à un traitement numérique qui utilise un mécanisme d'apprentissage basé sur l'algorithme Support Vector Machine (SVM). Ce mécanisme est effectué dans une optique empirique qui utilise la valeur des paramètres estimée à partir d'un corpus d'apprentissage étiqueté.
Rappelons que l'algorithme d'apprentissage SVM a déjà fait ses preuves dans plusieurs travaux de recherche dans le domaine de résumé automatique (e.g. : Chali et al. [Chali 2009] pour le résumé de documents multiples, Kianmehr et al. [Kianmehr 2009] pour le résumé de documents uniques, etc.).
Cet apprentissage permet de déterminer parmi les phrases non porteuses d'information rhétorique et ayant des relations "Autres - آخر" (i.e. la relation rhétorique "Autres - آخر" est attribuée lorsqu'aucune relation rhétorique n'est déterminée) celles qui sont pertinentes pour l'extrait final.
L'approche proposée pour la génération d'extraits se compose de quatre étapes (voir figure 5.1) à savoir : la segmentation du texte source en plusieurs unités textuelles plus petites, l'étiquetage morphologique des unités textuelles segmentées, l'analyse rhétorique, la sélection et le classe-

ment des unités textuelles pertinentes selon un profil utilisateur.
Chacune de ces étapes, comme indiqué plus haut, repose soit sur une analyse symbolique, soit sur un traitement numérique.
Ce procédé permet de coupler les informations linguistiques avec les traitements numériques afin de profiter des apports et avantages de ces deux aspects dans l'amélioration de la qualité du résumé généré.
Une des originalités de l'approche proposée est sa capacité d'identifier des relations rhétoriques précises, adaptées au type de résumé choisi et grâce auxquelles la structuration du texte source et l'identification des phrases pertinentes pourront être effectuées.

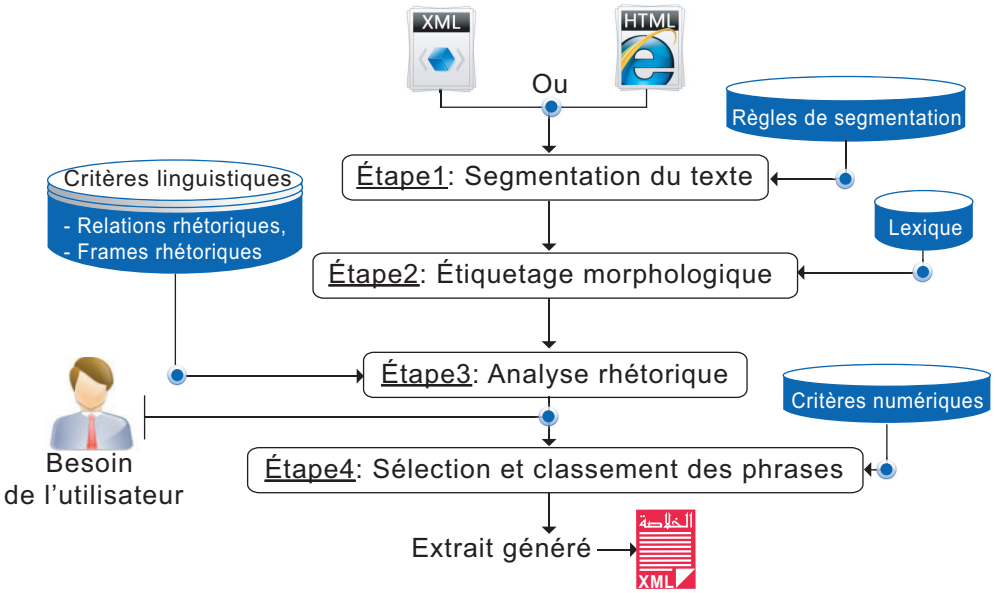

FIGURE 5.1 – Étapes de l'approche hybride proposée [Maaloul 2011]

5.2.1 Segmentation du document source

Un résumé de type extrait, est fondé sur l'hypothèse qu'il existe, dans tout texte, des unités textuelles saillantes [Amardeilh 2005], ce qui implique que l'étape de segmentation est incontournable pour la détection des unités textuelles saillantes.
Les unités textuelles considérées sont en général la phrase, ou un ensemble de phrases liées entre elles par des liaisons discursives, ou encore paragraphes. En effet, l'objectif de cette étape consiste à hiérarchiser et à structurer le texte source (qui se présente en format HTML / XML) en différentes unités textuelles plus petites : titres, sections, paragraphes et phrases.
Puisque les documents de notre corpus sont de type HTML ou XML, notre méthode de segmentation prend en considération les balises HTML de mise en forme qui mettent en relief certains passages.

Le tableau 5.1 illustre la liste des balises HTML de mise en forme utilisées dans le corpus afin de marquer le début et la fin des passages du texte. Ceci rend les textes structurés.

TABLEAU 5.1 – Liste des balises HTML de mise en forme utilisées pour la détection des passages début et fin

Balise HTML	**Type du passage délimiter**
<P> et </P> <DIV> et </DIV> <SPAN> et </SPAN>	Début et fin de paragraphe
 et 	Début et fin de titre
 et 	Début et fin de sous-titre

La méthode que nous proposons pour désambiguïser les frontières des phrases et des paragraphes, consiste à utiliser un ensemble de règles de segmentation défini par Belguith et al. [Belguith 2005] et des balises HTML de mise en forme utilisées dans le corpus afin de marquer le début et la fin des passages du texte.
Les règles de segmentation utilisées par Belguith et al. pour les textes arabes, se basent principalement sur 183 règles de segmentation en phrases et propositions. Ces règles reposent sur une méthode d'exploration contextuelle des signes de ponctuation, des conjonctions de coordination et de certains mots outils jouant le rôle de séparateur de phrases [Belguith 2008].

5.2.2 Étiquetage morphologique

L'objectif principal de l'étape d'analyse morphologique consiste à déterminer une représentation morphologique pour chaque mot constituant un texte arabe non voyellé. Nous proposons alors dans cette étape, d'utiliser la méthode proposée par Belguith [Belguith 1999].
Cette méthode d'étiquetage morpho-syntaxique permet de générer, pour chaque mot :

i sa *catégorie grammaticale* (i.e. : nom propre, adjectif, verbe, pronom possessif, pronom relatif, pronom démonstratif, préposition, conjonction de coordination, etc.) et,

ii ses *caractéristiques morphologiques* à savoir : son genre, son nombre, son temps, sa personne, sa détermination (déterminé/ non-déterminé) et son trait sémantique (humain/

non-humain).

Le principe de base de cette méthode d'étiquetage consiste à compenser la perte d'informations due à l'absence des signes de voyelles par un enrichissement du lexique de base, initialement réduit aux racines et aux affixes du langage, par des Formes Canoniques (FC) non verbales (i.e., la forme du masculin-singulier ou bien la forme du féminin-singulier en cas où cette dernière n'est pas dérivée de la première) associées aux mots ambigus du vocabulaire [1]. Cet enrichissement est utilisé comme moyen de filtrage des listes de caractéristiques présumées du mot à étiqueter (voir section section 3.4.4 du chapitre 2).
Comme nous l'avons mentionné dans la même section 3.4.4 du chapitre 2, la base de connaissance lexicale est formée par 3266 racines trilitères et quadrilitères de la langue arabe.
Cependant, chaque mot est préalablement décomposé en racine et affixes afin de fournir ultérieurement une liste de ses caractéristiques morpho-syntaxiques possibles (catégorie, type, temps, etc.).

5.2.3 Analyse rhétorique

L'étape de l'analyse rhétorique a pour but de comprendre le texte. Ainsi, son objet est de fournir un cadre d'interprétation pour la structure discursive du texte source d'une part et d'autre part de le d'écrire sous une représentation hiérarchisée qui met en évidence les différentes structures visuelles du texte et les propriétés qu'elles entretiennent entre elles.
Cette étape utilise le principe de la Théorie des Structures Rhétoriques définie par les chercheurs Mann et Thompson [Mann 1988] afin de générer une articulation cohérente entre les différentes unités minimales du texte source.
En effet, plusieurs frames rhétoriques, d'ordre linguistique et morphologique, ont été utilisés pour i) relier deux unités minimales adjacentes entre elles dont l'une posséde le statut de noyau – segment de texte primordial pour la cohérence – et l'autre ayant le statut noyau ou satellite – segment optionnel, et ii) déterminer les relations rhétoriques qui existent entre les différentes unités minimales juxtaposées, d'un même segment.
Outre l'utilisation classique de la technique RST, cette étape fait appel à deux phases à savoir la phase d'enrichissement et la phase de correction des relations rhétoriques pour procéder à la construction d'un seul arbre RST au-lieu de plusieurs arbres.
Dans les acquis, on signalera tout d'abord l'apport des frames rhétoriques basés sur des critères morphologiques dans l'enrichissement et la résolution de l'ambiguïté des relations rhétoriques détectées préalablement suite à des frames rhétoriques basés sur des critères purement linguistiques.

1. Dans le lexique des FC, les entrées sont classées selon le schème du pluriel, son type, et le / les schème(s) du singulier correspondant(s).

Un autre acquis est que cette première phase fait appel à une deuxième phase qui corrige et choisit les relations de discours les plus idéales en se basant sur des règles de correction.
Rappelons également que la relation rhétorique "Autres - آخر" est attribuée lorsqu'aucune relation rhétorique n'est déterminée à partir des frames rhétoriques (basés sur des critères linguistiques et des critères morphologiques).
Finalement, la construction de la structure globale d'un texte (arbre RST) consiste en une série de regroupements hiérarchiques de segments textuels effectués par l'intermédiaire de schémas rhétoriques et de quelques frames rhétoriques qui recouvrent récursivement les unités du texte. Ainsi, chaque schéma est défini par une relation rhétorique (e.g. : "Explication - تفسير", "Affirmation - جزم", "Explication - تفسير", etc.) qui spécifie sous quelles conditions deux segments peuvent être regroupés pour former un segment de niveau supérieur [Maaloul 2010b].

5.2.4 Phase d'apprentissage

En plus de l'analyse symbolique (segmentation, étiquetage morphologique et analyse rhétorique), le processus d'extraction des phrases pertinentes repose aussi sur l'apprentissage.
Généralement l'apprentissage, dans un contexte de résumé automatique, est utilisé pour déterminer l'importance d'une phrase à travers l'analyse préalable d'un corpus initial constitué par des collections de documents sources accompagnés de leurs résumés de référence [Jaoua 2011].
Nous proposons d'utiliser dans cette phase, l'apprentissage afin d'identifier des phrases saillantes et candidates pour l'extrait final (voir figure 5.2).
Dans ce cadre, le but est d'appliquer une équation hyperplan (frontière de décision) entre les classes {pertinentes} et {non-pertinentes}, sur les unités textuelles non porteuses d'informations rhétoriques dans un premier temps, et de déduire une décision d'appartenance à l'extrait final pour ces dernières en deuxième temps.
Le principe de déduction de la décision d'appartenance à un extrait final, se fait à l'aide d'un score calculé en fonction des critères positionnels, lexicaux et de cohésion appliqués sur les phrases ayant des relations rhétoriques "Autres - آخر". Signalons que les critères adoptés pour la sélection des phrases saillantes et candidates pour l'extrait final se basent sur :
i) une simplification de l'arbre RST, qui prendra en considération le type de résumé ou la liste des relations retenues par l'utilisateur, et
ii) un taux de réduction qui favorise les phrases les plus pertinentes [2].
En s'appuyant sur ce principe (voir figure 5.4), une phrase est considérée plus pertinente lorsque son point :
a- se trouve dans une région pertinente, et
b- possède la distance la plus loin de tous les autres voisins par rapport au classificateur linéaire

2. En utilisant l'algorithme SVM, chaque phrase est définie par un point.

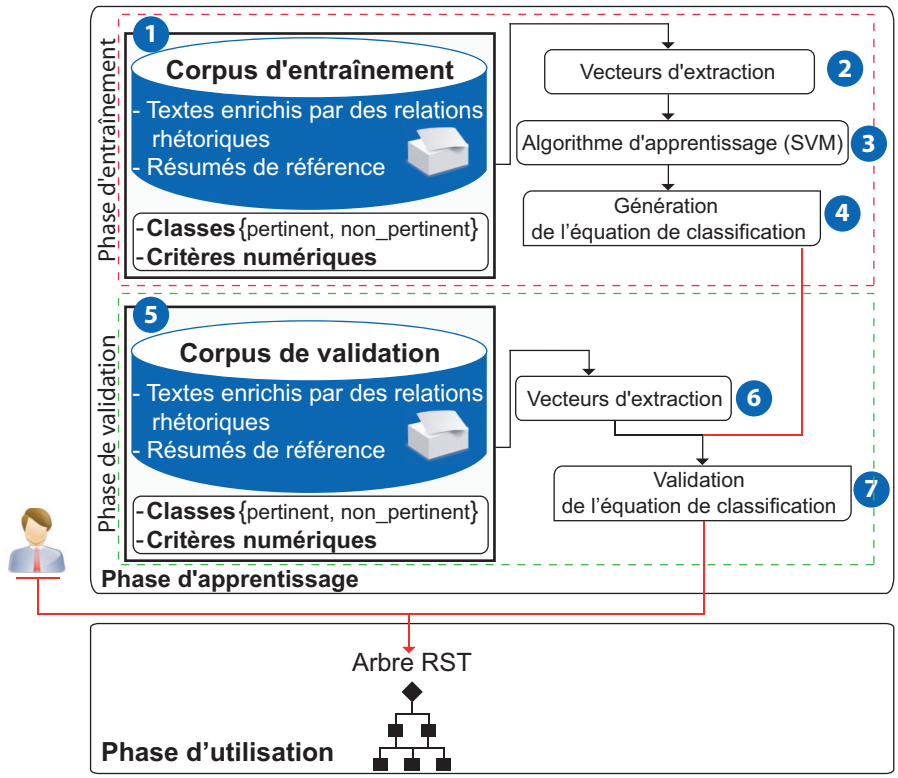

FIGURE 5.2 – Principe d'apprentissage

appelé hyperplan.

5.2.4.1 Corpus d'apprentissage

La construction d'un corpus de référence [3] n'est pas un problème propre au résumé automatique. C'est en soi un problème auquel se trouve confronté l'ensemble de la communauté de TALN [Minel 2002a]. Ainsi, les problèmes du genre textuel, d'homogénéité, de représentativité de certains phénomènes linguistiques surgissent immédiatement.

Les textes que nous avons utilisés dans la phase d'apprentissage ont été sélectionnés à partir du corpus de la magazine "Dar al Hayat - "دَار الحيَاة" [4]. Le choix du domaine journalistique se justifie principalement par le nombre, généralement important, de marqueurs linguistiques [Maâloul 2008].

On notera aussi, que certains genres de textes sont plus fréquents que d'autres ; ce choix est motivé parce que certains genres représentent une certaine diversité et richesse des événements et des expressions temporelles (par exemple : l'actualité politique) par rapport aux autres genres (e.g. : science).

3. corpus d'apprentissage

4. Dar Al Hayat ('Maison La Vie') est un journal quotidien généraliste arabophone avec une diffusion dans le monde arabe de 110 000 exemplaires

Les expériences d'apprentissage sont faites sur un corpus d'apprentissage composé de 185 paires {document et résumé} d'articles journalistiques venant de 5 genres, avec une moyenne de 33 phrases par document.
La distribution des textes choisis en fonction de leur genre est résumée dans le Tableau suivant :

TABLEAU 5.2 – Caractéristiques du corpus d'apprentissage

Genre	**Nombre de documents**	**Taille des documents (en nombre de phrases)**
Actualité sportive	25	200
Actualité nationale	45	1800
Politique	50	2800
Éducation	30	660
Science	35	490
Total	185 documents	5950 phrases

Afin d'accélérer le processus d'apprentissage, nous avons opté à une étape d'étiquetage des textes, par deux experts, pour repérer la liste des phrases pertinentes qui exprime le mieux possible l'idée illustrée par le texte source.
L'étiquetage des textes sources a été effectuée entièrement à la main et la détermination du résumé de référence est déduite après un accord total entre les experts.
Il est à noter que le résumé de référence est constitué de phrases extraites du texte source.
Ci-après, nous présentons un exemple (voir figure 5.3) de texte étiqueté de notre corpus de référence. Notons que les phrases formant le résumé de référence sont encadrées par le caractère "$".

FIGURE 5.3 – Exemple de texte étiqueté : texte source avec son résumé de référence

5.2.4.2 Apprentissage basé sur l'algorithme SVM

Notre mécanisme d'utilisation de l'apprentissage se place dans un contexte bien défini : nous nous intéressons à la question de l'usage des SVM pour les phrases ayant des relations rhétoriques "Autres - آخر", afin de produire une déduction : les quelles de ces phrases sont pertinentes pour l'extrait final ?

Ainsi, l'utilisation d'une technique d'apprentissage est rendue possible suite à l'élaboration d'un corpus de référence formé de cinq mille neuf cent cinquante exemples de phrases annotées et relatives au cent quatre-vingt-cinq articles de presse. En effet, l'objectif de cette phase est d'apprendre à classifier en fonction des critères sur lesquels se sont appuyés les experts pour élaborer leurs résumés de référence.

À partir de notre corpus d'apprentissage composé de textes sources et de leurs extraits, une première étape consiste à appliquer les critères d'extraction des phrases. Ces critères combinent des informations d'ordre positionnel (e.g. : la phrase est située à la première section), lexical (la phrase contient des mots fréquents dans le texte), structurel (la phrase contient des mots présents dans le titre du texte), etc.

Dans une deuxième étape, chaque phrase du texte source [5] est analysée en fonction des critères d'extraction (voir annexe B), puis comparée à l'extrait de référence. Cette étape aboutit à la construction d'un ensemble de vecteurs d'extraction [6].

La troisième étape utilise l'algorithme d'apprentissage SVM pour produire un hyperplan optimal qui sépare deux classes (pertinente et non-pertinente), où les vecteurs d'extraction constituant l'entrée à partir desquelles l'équation d'hyperplan est construite (voir figure 5.4) .

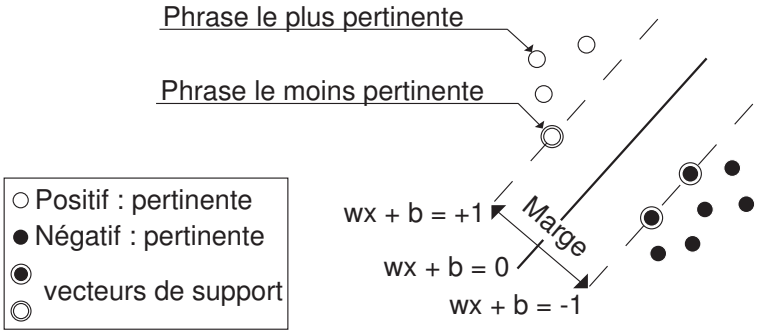

FIGURE 5.4 – Support Vector Machine

Pour cela, la détermination de l'équation de classification nécessite une base de données S [7] de

5. Document tiré de notre corpus de référence ou d'apprentissage.

6. Vecteur d'extraction : vecteur contenant des valeurs numériques et/ou des classes relatives aux critères d'extractions. Il contient également la classe d'appartenance de la phrase à l'extrait de référence {pertinente ou non-pertinente

7. S : L'ensemble des vecteurs d'extraction

n [8] points d'un espace de dimension P appartenant à deux classes différentes t_i qu'on notera la classe {-1} et la classe {+1}.

$$S = \{X_i, t_i | X_i \in \mathbb{R}^p; t_i \in \{-1, +1\}; i = 1...n\} \tag{5.1}$$

L'équation d'hyperplan optimal est déduite sous la forme suivante :

$$Score_{phrase} = \sum_{i=1}^{n} \alpha_i (X_i + b) \tag{5.2}$$

La dernière partie (phase d'utilisation) permet de produire une décision relative à chaque phrase du texte source et cela en comparant le $Score_{phrase}$ comme suit [Steinwart 2008] :

$$\begin{cases} \text{if } (Score_{phrase} > 0), & Classe\ de\ X_i = \{+1\} \\ \text{if } (Score_{phrase} < 0), & Classe\ de\ X_i = \{-1\} \end{cases} \tag{5.3}$$

5.2.4.3 Vecteur d'extraction SVM

L'acte annotatif consiste à donner une valeur ou un jugement à un passage d'un texte en se référant aux critères d'extraction. Un passage peut-être soit important soit non important [Li 2007]. Dans notre cas, un passage présente une unité du texte qui est une phrase.
Les exemples d'entraînement sont donnés sous la forme d'un vecteur d'extraction V, où chaque vecteur décrit une phrase sous la forme suivante :

$$V_i(S_1, S_2, S_3, ..., S_j, ...S_n, D_i) \tag{5.4}$$

Avec :

- V_i : vecteur d'extraction de la phrase i
- j : critère d'extraction numéro j
- n : nombre de critères d'extraction
- S_j : le score de la phrase i selon le critère d'extraction j
- D_i : critère booléen représentant la classe du vecteur V_i : {pertinent, non-pertinent}

En règle générale, tout vecteur V est décrit par la collection de valeur S_j présenté par un *score* et/ou une *classe*, où la valeur donnée d'un critère correspond à la valeur d'analyse de la phrase selon ce critère. Chaque vecteur est complété par un critère booléen de classe D_i. Ce critère de classe est déduit suite à une simple vérification de la présence ou l'absence de la phrase dans l'extrait de référence.
Toutefois, la répétition d'un certain nombre de vecteurs identiques, et la contradiction d'autres est possible.

8. n : Position d'une phrase dans l'espace de dimension P

5.2.4.4 L'algorithme SMO

L'utilisation des algorithmes SVMs d'apprentissage standards était limitée à un groupe de chercheurs car ces algorithmes standards étaient longs et difficiles à implémenter. Pour Palier à ce problème, John. C.Platt [Platt 1998] à mis au point un algorithme d'apprentissage appelé SMO [9] "*Sequential Minimal Optimization*" qui permet de résoudre rapidement le problème d'optimisation quadratique (QP). Cet algorithme est généralement plus rapide, plus simple à implémenter et nécessite un espace mémoire réduit [Keerthi 2001].
Suite au principe que la validation [10] est une phase indispensable à tout processus d'apprentissage, notre phase de validation a pour but de raffiner l'équation de classification générée par l'algorithme SMO.
Le processus de validation utilise un corpus de validation formé par 5 paires {document et résumé} d'articles journalistiques, pris du corpus de la magazine "Dar al Hayat - دَار الحيَاة, venant de 5 genres différents (Actualité sportive, Actualité nationale, Politique, Éducation et Science) avec une moyenne de 35 phrases par document, et se base sur un calcul comparatif des mesures de rappel, de précision et de F-mesure.
Le corpus de validation, malgré sa taille relativement petite, nous a permis d'avoir une idée sur le noyau "*kernel*" à choisir parmi la liste de noyau fournis par l'outil Weka [11] et qui correspond à l'équation optimale à utiliser dans la phase d'utilisation.
Ainsi, en comparant les mesures moyennes de rappel, de précision et de F-Mesure données par l'outil Weka après la phase d'apprentissage et de validation, nous avons constaté que le noyau polynomial [12] donne les meilleures valeurs pour ces mesures (ces mesures sont respectivement 71,7%, 51.5% et 59,9% alors que les mesures données par le noyau gaussian [13] sont respectivement de 66,7%, 48,5% et 56,1%). Nous remarquons, une légère amélioration donnée par le noyau polynomial, ce qui justifie le choix de ce dernier dans la phase d'apprentissage basée sur l'algorithme SMO.

9. L'idée principale de SMO consiste à optimiser seulement deux vecteurs par itération. En effet, l'algorithme SMO optimise la fonction objective duale du problème global en opérant à chaque itération un ensemble réduit à deux multiplicateurs de Lagrange. SMO permet, ainsi, de résoudre le problème de programmation quadratique sans nécessité de stocker une grande matrice en mémoire et sans une routine numérique itérative pour chaque sous problème. [Platt 1998]
10. La validation consiste à vérifier que le modèle construit sur l'ensemble d'apprentissage permet de classer tout individu avec le minimum d'erreurs possible.
11. Weka : `http://weka.wikispaces.com/` est une plateforme développée par Université de Waikato, contenant une collection d'outils de visualisation et d'algorithmes pour l'analyse des données et l'apprentissage automatique [Steinwart 2008].
12. Dans les paramètres de la méthode SMO, le noyau polynomial "PolyKernel" est utilisé par défaut.
13. Noyau gaussian noté dans la liste des paramètres de la méthode SMO dans l'outil Weka par "RBFKernel" [Witten 2005].

5.2.4.5 Scénario d'application de l'algorithme SVM

Le scénario du manque de l'information rhétorique après l'étape de l'analyse rhétorique, et le risque de la sélection des phrases moins ou même non pertinentes dans l'extrait final, sont les problèmes que nous voyons souvent apparaître lors de la génération de l'extrait final. Ainsi, par exemple, une phrase pertinente non retenue dans l'extrait final parce que, tout simplement, elle est non porteuse de relation rhétorique.
Le scénario du manque d'information linguistique peut être produit lorsqu'un segment contient une relation rhétorique "Autres - آخر" à cause par exemple de l'absence d'un indicateur déclencheur de relation rhétorique.
Dans la même logique, si ce type de phrase est susceptible d'être non retenue, on risque d'avoir dans l'extrait juste les phrases porteuses des relations rhétoriques et qui peuvent êtres moins pertinentes que d'autres non porteuses d'informations linguistiques.
Pour résoudre ce problème, nous proposons d'utiliser un traitement numérique afin de faire une décision quant aux phrases non porteuses d'informations rhétoriques.

5.2.5 Sélection et classement des phrases selon le type du résumé

Suite au principe, qu'une information n'est pas importante en soi, mais doit correspondre aux besoins d'un utilisateur [Maâloul 2008], notre approche hybride aborde la question des besoins des utilisateurs. Ainsi, il s'agit dans cette étape de classer et de sélectionner à partir du texte source un sous ensemble de phrases répondant à un certain nombre de critères liés au type du résumé choisi par l'utilisateur et au scénario d'application de l'algorithme SVM.
Toutefois, cette étape consiste d'abord à classer les phrases selon le type du résumé choisi par l'utilisateur (informatif, indicatif ou d'opinion), et ensuite à faire une sélection des phrases les plus importantes dans l'arbre RST afin de les afficher dans l'extrait final.
Suite à ce principe, notre recherche s'oriente alors vers la production de résumé dynamique.

5.2.5.1 Classement et sélection des phrases

Pour le résumé, ce ne sont pas toutes les phrases ayant des unités noyaux qui sont considérées comme importantes. En effet, cette étape, consiste à classer des phrases contenant les unités minimales importantes - noyaux, et cela en profitant des relations entre les structures de discours pour en décider du degré de leur importance.
Le classement des phrases de l'arbre RST prendra en considération le type du résumé adapté pour la génération du résumé final.
Notons que notre proposition cible les besoins potentiels d'un utilisateur. Nous offrons alors, à l'utilisateur, la possibilité de construire ses propres itinéraires à travers le texte et ce en choisissant le type du résumé qui l'intéresse.

Il s'agit d'un résumé dynamique qui peut être généré en fonction des intérêts et du profil de l'utilisateur (résumé indicatif, informatif, etc.).
En cas où ce dernier ne précise aucun choix, le système détermine automatiquement les relations retenues pour le type de résumé indicatif. En effet, le type de résumé indique les relations rhétoriques adéquates pour la génération du résumé.
Ainsi, suite à l'étude analytique menée sur une centaine de résumés, réalisés par deux experts sur les documents de notre corpus, nous avons remarqué que généralement, un résumé indicatif, un résumé informatif et un résumé d'opinion sont déterminés par la liste des relations rhétoriques suivantes :

TABLEAU 5.3 – Liste des relations rhétoriques retenues pour le type de résumé indicatif, informatif et opinion

Type de résumé	**Liste des relations rhétoriques**
Résumé indicatif	"Confirmation - توكيد", "Conclusion - اسـتنتَاج" , "Affirmation - جزم"
Résumé informatif	"Confirmation - توكيد", "Conclusion - اسـتنتَاج" , "Affirmation - جزم", "Justification - تعليل", "Réduction - تقليل", "Énumération - تفصيل", "Pondération - ترجيح", "Possibilité - إمكَان"
Résumé d'opinion	"Condition - شرط", "Concession - استدرَاك", "Exception - اسـتثنَاء", "Confirmation - توكيد", "Évidence - قَاعدة", "Négation - نفي", "Conclusion - اسـتنتَاج", "Affirmation - جزم", "Pondération - ترجيح", "Possibilité - إمكَان", "Restriction - حصر", "Justification - تعليل"

5.2.5.2 Vers un résumé dynamique

L'un des objectifs que nous avons ciblé en proposant une approche hybride pour le résumé automatique de documents arabes est la prise en compte des besoins potentiels d'un utilisateur. Ainsi, l'utilisateur a la possibilité de lancer un traitement de résumé automatique sur un article de presse, tout en intégrant des paramètres qui vont conditionner le futur résumé, tels que le type de résumé (indicatif, informatif, d'opinion, etc.) ou la liste des relations rhétoriques, le taux de réduction mentionné en nombre de phrases, et en choisissant une stratégie de sélection à suivre (l'application ou non de la phase d'apprentissage).
Signalons que chaque type de résumé est déterminé par une liste de relations rhétoriques. Dans le tableau 5.3 nous avons mentionné la liste des relations rhétoriques retenues pour les résumés de type indicatif, informatif et opinion.
Rappelons, que la définition de la liste des relations formant un type de résumé est déduite

suite à une étude analytique et après un accord total mené sur une centaine de résumés réalisés par deux experts. La réduction de l'arbre RST se fait par la suppression de toutes les phrases qui viennent d'une relation rhétorique non retenue pour un type de résumé.
Pour le taux réduction, qui représente le nombre de phrases extraites par rapport au nombre de phrases contenues dans le document original, notre approche favorise les phrases les plus pesantes selon un score défini à partir de l'étape d'apprentissage. Cette contrainte consiste à favoriser la phrase la plus éloignée à l'hyperplan par rapport à d'autres qui sont plus proches (voir figure 5.4).
Ainsi, l'extrait final ne gardera que les phrases qui respectent premièrement le type du résumé ou l'ensemble des relations rhétoriques sélectionnées, et deuxièmement celles qui ont le meilleur score qui reflète la distante d'éloignement par rapport aux frontières de décision du coté pertinent.
Par ailleurs, en cas d'égalité entre les scores des phrases, notre approche favorise les phrases selon leurs positions d'apparition dans le document source, et selon leurs tailles (favorisant les phrases les plus courtes). L'utilisateur a ainsi la possibilité de construire un résumé adapté à ses besoins et d'une façon dynamique.

5.3 Conclusion

Dans le présent chapitre, nous avons présenté les principales étapes de notre approche d'extraction qui opère selon une vision hybride. Nous avons, ainsi, proposé deux étapes de prétraitement de documents (segmentation et étiquetage morphologique) et trois autres étapes pour l'analyse selon la vision hybride (analyse rhétorique, apprentissage, classement et sélection des phrases).
Au niveau de l'apprentissage, nous avons proposé l'utilisation de l'algorithme d'apprentissage SVM en vue de produire une décision sur l'importance d'une phrase non porteuse d'informations rhétoriques. Cette décision se base sur un score déterminé à partir d'un vecteur d'extraction et un hyperplan optimal qui sépare deux classes (pertinente et non-pertinente).
Au niveau du classement et sélection des phrases saillantes, nous avons proposé de tenir en couple les besoins de l'utilisateur. En effet, nous considérons qu'une information n'est pas importante en soi, mais doit correspondre aux besoins d'un utilisateur. Ainsi, nous avons proposé de classer et de sélectionner un sous ensemble de phrases répondant à un certain nombre de critères liés au type du résumé choisi par l'utilisateur et au scénario d'application de l'algorithme SVM.
Le chapitre suivant décrit la mise en œuvre des concepts que nous avons décrits tout au long des chapitres 4 et 5. Nous mentionnons, principalement, les structures et modules que nous avons intégrés dans notre système de génération d'extrait.

Troisième partie

Un système hybride de résumé automatique

Chapitre 6

Architecture et description du système

Sommaire

6.1 Introduction 95
6.2 Architecture du système 96
6.3 Implémentation : du texte brut à l'extrait 97
6.3.1 Chargement du texte 98
6.3.2 Segmentation du texte 99
6.3.3 Étiquetage morphologique 100
6.3.4 Analyse rhétorique du texte 102
6.3.5 Détermination de l'arbre RST 103
6.3.6 Sélection des critères d'extraction reflétant les besoins de l'utilisateur . . . 104
6.3.7 Sorties du système 106
6.4 Conclusion 107

6.1 Introduction

L'approche de génération d'extrait, que nous avons proposée tout au long des chapitres 4 et 5 se distingue par sa vision hybride qui permet la prise en compte des besoins potentiels d'un utilisateur ; ce qui permet de converger vers un extrait dynamique. Cette vision offre une nouvelle démarche qui s'articule autour de la génération d'extrait par instanciation d'une approche basée sur l'utilisation conjointe d'une méthode symbolique et numérique.
L'objectif, est de profiter de l'avantage des résultats explicites fournis par la théorie de la structure rhétorique (RST) et de l'apport de la technique d'apprentissage basée sur l'algorithme Support Vector Machine (SVM), grâce auquel la structuration du texte source et l'identification des phrases pertinentes pour l'extrait final pourront être effectués selon un type de résumé choisi.
Dans ce chapitre, nous abordons dans un premier temps l'architecture générale de notre système "L.A.E (**L**axAS **A**l **E**l|ly) - الـلَخَاص الآلي", une implémentation informatique pour la génération d'un extrait, tout en détaillant les modules principaux permettant de le mettre en œuvre. Nous décrirons, en deuxième temps, quelques aspects d'utilisation du système.

6.2 Architecture du système

Le schéma synoptique du système "L.A.E (**L**axAS **A**l **E**l|ly) - الـلَخَاص الآلي" est composé de trois principaux modules dont l'interaction est détaillée dans la figure 6.1.

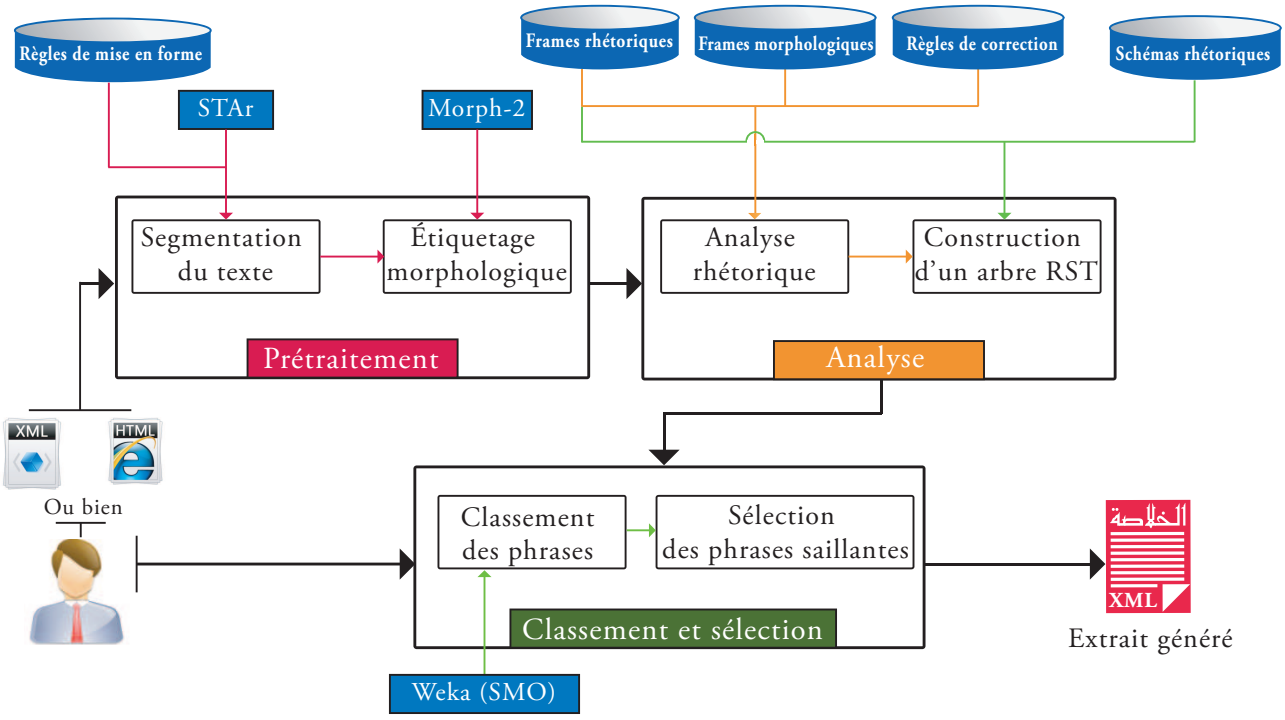

FIGURE 6.1 – Architecture du système "L.A.E - الـلَخَاص الآلي"

L'architecture du système "L.A.E - الـلَخَاص الآلي" distingue les modules suivants :

- Le module de pré-traitement : ce module permet de segmenter un fichier HTML ou XML selon plusieurs niveaux (e.g. titres, sous titres, sections, paragraphes, phrases, etc.), et d'étiqueter morphologiquement tous les mots en déterminant la liste de toutes leurs caractéristiques morphologiques possibles. Ainsi, le module de pré-traitement se base en partie, sur l'utilisation d'un certain nombre d'outils. Notre réalisation à ce niveau exploite le texte source segmenté et balisé délivré par l'outil **STAr** [Belguith 2005] dans un format XML afin de l'enrichir par des étiquettes morphologiques déterminées par l'outil **Morph-2** [Chaaben 2010].
- Le module d'analyse : il permet de détecter les relations rhétoriques qui relient deux unités minimales adjacentes entre elles. Ce module délivre un arbre RST qui qualifie les relations détectées entre les unités textuelles. Les relations rhétoriques de l'arbre RST issus de ce module permettent de participer avec certains critères dans la génération de l'extrait final.

– Le module de classement et sélection : il permet de classer les phrases de l'arbre RST selon un type de résumé ou une liste de relations rhétoriques choisies par l'utilisateur. Il assure, en outre, la sélection des phrases les plus pertinentes en fonction des critères hybrides qui ciblent les besoins d'un utilisateur (e.g. type de résumé, taux de réduction, etc.).

Ainsi, le système "L.A.E - الـلَخَاص الآلي" produit à la fin un résumé (sous format XML) à partir de l'arbre RST et en fonction des besoins de l'utilisateur. Il est à signaler que l'implémentation est faite avec le langage de programmation Delphi et la modélisation est réalisée avec le langage UML.
Nous avons choisi le standard XML pour l'échange d'information entre les différents modules du système.
Dans la suite de cette section, nous détaillons les différents modules du système "L.A.E - الـلَخَاص الآلي" que nous venons d'énumérer.

6.3 Implémentation : du texte brut à l'extrait

Nous avons développé une interface graphique principale en Delphi gérant en amont tous les traitements nécessaires aux tâches demandées par l'utilisateur afin de générer un extrait relatif à son besoin. Cette interface regroupe les traitements suivants que nous présenterons dans la suite : segmentation, étiquetage morphologique, analyse rhétorique, détermination de l'arbre RST, apprentissage basé sur l'algorithme SVM et génération du résumé.
Dans cette interface principale, le système présente à l'utilisateur les différentes étapes d'analyse du texte.
Rappelons que notre contribution se base principalement sur une interaction avec l'utilisateur. Par ailleurs, notre objectif principal est de sélectionner des unités textuelles (principalement des phrases) qui correspondent au profil de filtrage fixé par l'utilisateur. Le simple fait de modifier ce profil permet d'obtenir un extrait différent.
La plate-forme logicielle du système, étant donné, son caractère évolutif, donne la possibilité d'enrichir et de manipuler diverses tâches selon une organisation des marqueurs, des indices nécessaires, des relations rhétoriques, etc.
Pour ce faire, nous pouvons transposer les connaissances linguistiques (mises dans des frames rhétoriques) dans le modèle conceptuel d'exploration contextuelle en spécifiant les indicateurs, les indices, les parties noyaux et/ou satellites et les règles d'exploration contextuelle associées.
Pour effectuer la génération d'extrait, l'utilisateur n'a qu'à sélectionner le texte qu'il désire annoter et suivre les étapes décrites dans la zone cadrée en rouge à gauche de la figure 6.2.

FIGURE 6.2 – Interface principale du système "L.A.E - الـلّـخَاص الآلي"

6.3.1 Chargement du texte

L'étape de chargement d'un texte nécessite deux étapes à savoir l'identification de la source du texte[1] et sa sélection (voir figure 6.3). En effet, le système traite deux formats de fichiers (HTML et XML).

Rappelons que l'identification de la source du texte, en particulier de l'article de presse permet de guider son processus de segmentation et ce en déterminant les balises HTML de mise en forme relatives à ce texte.

Par exemple, pour les articles de presse, que nous avons rapatriés du journal électronique "Dar al hayat - دَار الحيَاة", nous avons remarqué que les paragraphes sont délimités par la balise $< P >$ et $< /P >$, le titre par les balises $< fontstyle = "font - family : SimplifiedArabic, ABGeeza, TimesNewRoman, Times; color : Black; font - size : 15pt;" >et< /font >$, et les sous-titres par les balises $< fontstyle = "font - family : SimplifiedArabic, ABGeeza, TimesNewRoman, Times; color : Black; font - size : 13pt;" >$ et $< /font >$.

1. Journal : Dar el Hayet (http://http://www.daralhayat.com/), Al jazerra (http://http://http://www.aljazeera.net/), Annahar (http://www.annahar.com/), etc.

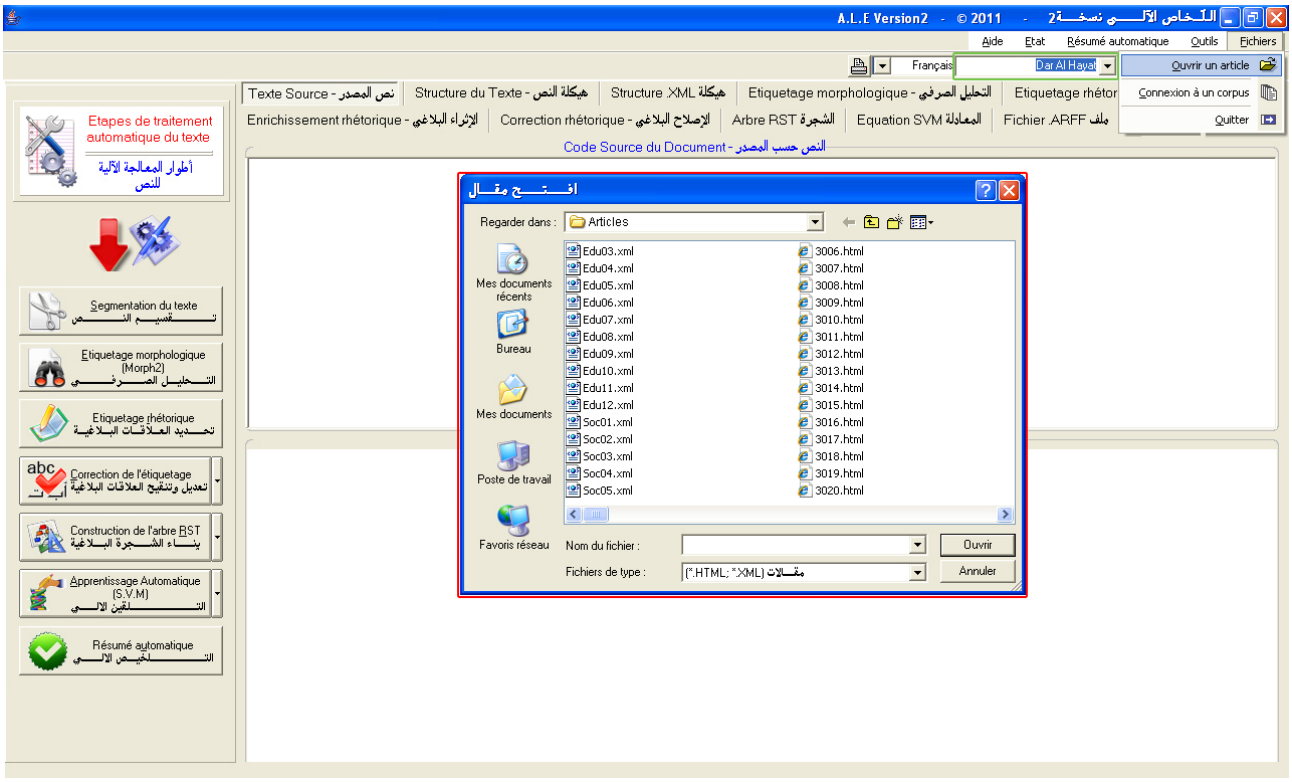

FIGURE 6.3 – Interface de chargement du texte

6.3.2 Segmentation du texte

Le module de segmentation du texte consiste à segmenter le texte source selon plusieurs niveaux : titre, sous-titres, paragraphes et phrases. (voir figure 6.4).

Ce module préconise, en entrée, une base de données présentée dans un fichier XML contenant les règles de segmentation basées sur des balises de mise en forme (qui mettent en relief certains passages) et par la suite délivre les paragraphes détectés à l'outil **STAr**. De son coté, STAr segmente le paragraphe entré en une liste de phrases.

Rappelons que la segmentation en phrases par l'outil STAr est basée sur l'exploration contextuelle des signes de ponctuation, des mots connecteurs jouant le rôle de séparateur de phrases (e.g., "lakin | لكن", "laqad | لقد" et "amma | أمَّا") ainsi que celle de certaines particules telles que les conjonctions de coordination ("wa | و" et "f | ف") [Belguith 2005].

Le résultat de ce module est mis en évidence par un document structuré en format XML contenant le titre, les sous-titres et les phrases du texte.

FIGURE 6.4 – Segmentation du texte en titres, sous-titres et phrases.

La figure 6.5 présente un exemple de résultat délivré par ce module.

6.3.3 Étiquetage morphologique

Le deuxième module **Morph-2** reçoit le fichier XML généré par le premier module (voir figure 6.5) et applique une analyse morphologique sur tous les mots des phrases mentionnées. Le module Morph-2 permet, à partir d'une analyse morphologique approfondie, de déterminer pour chaque mot non seulement sa racine et ses affixes mais aussi la liste de toutes ses caractéristiques morphologiques possibles, tout en tenant compte des différents types d'ambiguïtés cités dans 3.2 (l'agglutination, l'affixation, la transformation, etc.).

Morph-2, est un analyseur morphologique de textes arabes non voyellés basé sur cinq étapes à savoir, la segmentation du texte en mots, le pré-traitement morphologique, l'analyse affixale, l'analyse morphologique et le post-traitement [Belguith 2006].

Dans chaque étape d'analyse, Morph-2 utilise un ensemble de données nécessaires au traitement telles que les espaces, les signes de ponctuation et certains caractères spéciaux dans l'étape de *segmentation*, le lexique des proclitiques, enclitiques, et les particules dans l'étape de *pré-traitement morphologique*, le lexique des triades affixales et des racines (3266 racines trilitères et quadrilitères) dans l'étape d'*analyse affixale*, le lexique des noms dérivés et primitifs et le

```
<?xml version="1.0" encoding="utf-8" ?>
- <texte>
<titre> بروفة ليبية للخروج من سجن ٢٤ عاماً</titre>
- <section>
- <paragraphe>
  <phrase>حسنٌ أن ليبيا تتحرر من حكم الجاهلية العظمى بالقطعة</phrase>
  <phrase>وإلا فلا يمكن تخيل ما يحدثه التحرير دفعة واحدة من فوضى عارمة</phrase>
  <phrase>فليبيا بلاد كبيرة</phrase>
  <phrase>.وهذا قد يُصعب وضع كل شيء في مكانه الجديد بيسر وبلا أخطاء أو خسائر</phrase>
  <phrase>قطعاً لم يكن ذلك بتدبيرنا</phrase>
  <phrase>إنما بتدبير إلى هنا</phrase>
  <phrase>الذي جعل المدن سابقة ولاحقة في التحرر</phrase>
  <phrase>.علّنا نجتنب تكرار الأخطاء</phrase>
</paragraphe>
- <paragraphe>
  <phrase>واللافت في ثورتنا أن النظام السابق مازال طليقاً حتى الآن في بعض الأرجاء من بلادنا</phrase>
  <phrase>والإنسان لدينا نصفه مقاتل ونصفه ينتظر القتال</phrase>
  <phrase>.وذلك ما يضعنا في مسؤوليات أكبر تمتحن مدى قدرتنا على الاستقرار السريع</phrase>
</paragraphe>
+ <paragraphe>
+ <paragraphe>
+ <paragraphe>
+ <paragraphe>
+ <paragraphe>
+ <paragraphe>
</section>
- <section>
  - <soustitre>أحداث متلاحقة<soustitre>
    - <paragraphe>
      <phrase>هذا ملخص حالتنا العقلية</phrase>
      <phrase>فما المتوقع من شعب سُجن ٢٤ عاماً وتحولت أول بروفة في حياته على التظاهر إلى ثورة حقيقية</phrase>
      <phrase>ثم وهو في قلب حقيقة لا يكاد يصدقها</phrase>
      <phrase>...وجد نفسه مضطراً لتسليحها</phrase>
      <phrase>فما المتوقع من شعب سُجن ٢٤ عاماً وتحولت أول بروفة في حياته على التظاهر إلى ثورة حقيقية</phrase>
      <phrase>ثم وهو في قلب حقيقة لا يكاد يصدقها</phrase>
      <phrase> هل سيتوقف ويعود فوراً الى كهفه القديم كي لا تتحول الثورة إلى حرب ويستمر في مزاولة خوفه أم سيسير إلى النهاية مهما
      كانت الخسائر</phrase>
      <phrase>وهل ثمة حقاً ما يستحق العودة إلى وراء كان يخسر فيه كل شيء من دون أن يهم بفعل شيء</phrase>
    </paragraphe>
    + <paragraphe>
    + <paragraphe>
    + <paragraphe>
    + <paragraphe>
    + <paragraphe>
    + <paragraphe>
</section>
+ <section>
</texte>
```

FIGURE 6.5 – Exemple de fichier XML délivré par le module de segmentation

lexique des correspondances entre formes canoniques et formes dérivées dans l'étape d'*analyse morphologique*.

Soulignons que l'intégration de Morph-2 dans le système a nécessité une adaptation au niveau des étapes de segmentation et de post-traitement.

Ce module délivre comme résultat final un fichier XML. La figure 6.6 montre le résultat de l'analyse morphologique d'un texte enrichi par les balises de segmentation en phrases. Ainsi, pour chaque mot, on obtient ses différentes caractéristiques morphologiques possibles.

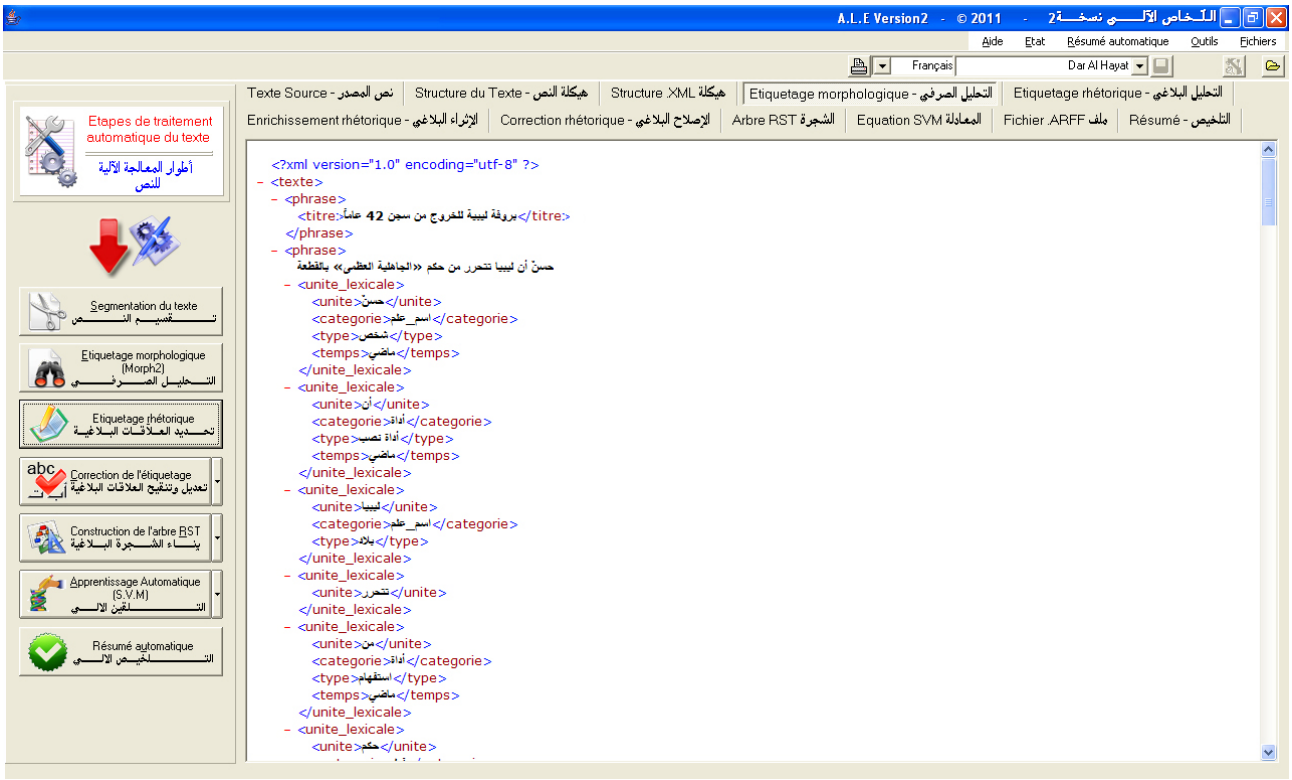

FIGURE 6.6 – Étiquetage morphologique du texte

6.3.4 Analyse rhétorique du texte

Le troisième module se charge de détecter les relations rhétoriques à partir des informations prises du fichier XML qui vient d'être généré par le deuxième module. Comme il a été décrit dans la section 4.3 du chapitre 4, la déterminations des relations rhétoriques se base sur trois étapes :

1. Relier deux unités minimales adjacentes entre elles (dont l'une possédant le statut de noyau et l'autre ayant le statut noyau ou satellite) et déterminer les relations rhétoriques qui existent entre les différentes unités minimales juxtaposées.

2. Enrichir cette liste de relations rhétoriques par d'autres relations détectées à l'aide des informations morphologiques.

3. Corriger les relations rhétoriques détectées dans les étapes 1. et 2. par d'autres, en se basant sur des règles de correction de types : relation-relation et/ou relation-indice.

Le résultat du module d'analyse rhétorique est un fichier XML qui présente la totalité des phrases du texte avec les relations qui les relient.

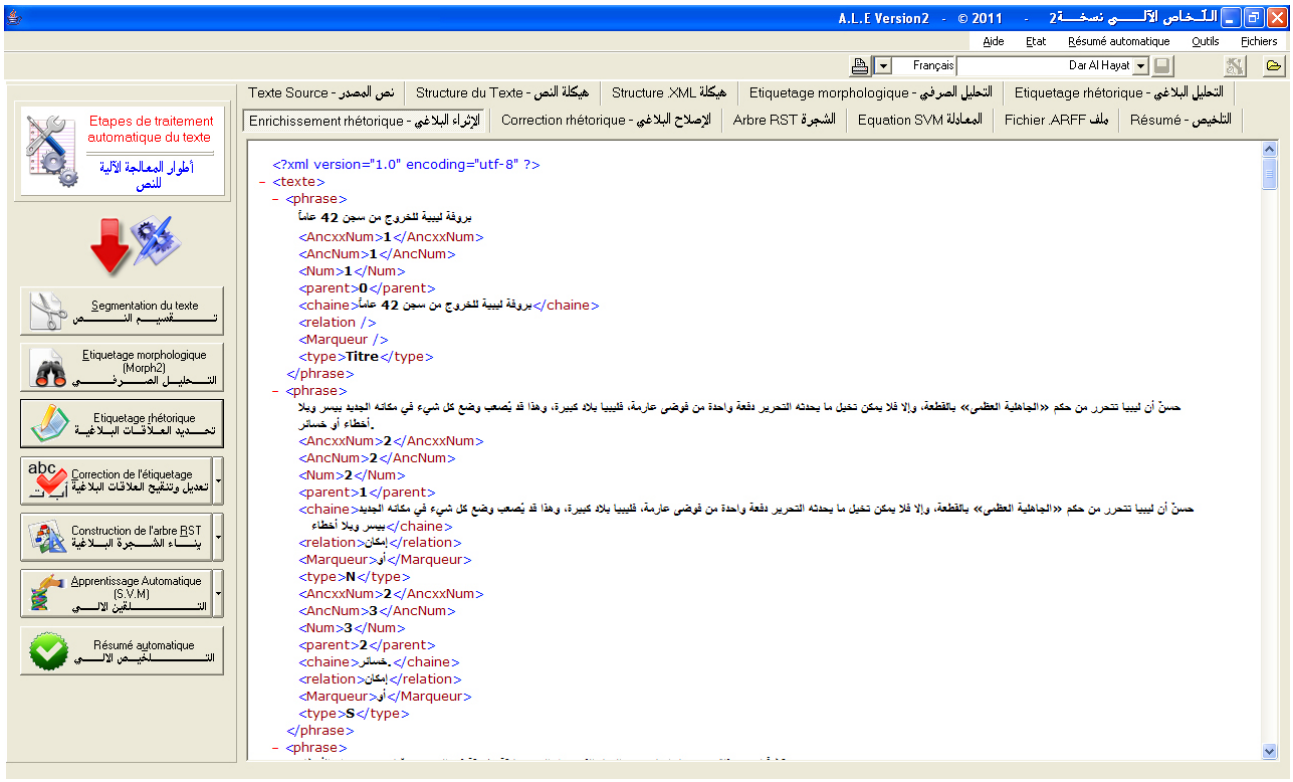

FIGURE 6.7 – Analyse rhétorique du texte

6.3.5 Détermination de l'arbre RST

Le module de détermination de l'arbre RST a pour but de construire l'arbre jugé le plus descriptif du texte et cela en utilisant cinq schémas définis en terme de relations (par Mann et Thompson [Mann 1988]) et quelques frames rhétoriques. Ces critères décrivent ainsi, l'organisation structurelle d'un texte, quel que soit le niveau hiérarchique de ce dernier.
Le module de détermination de l'arbre RST est appliqué selon un ordre cohérent et informatif, en amont de la liste des relations rhétoriques déterminées et qui lient des segments de texte adjacents.

Il est à souligner qu'un utilisateur peut accéder à la structure hiérarchique et cela en ajoutant, modifiant ou même supprimant une ou plusieurs relations. Il devient ainsi, possible de donner une nouvelle hiérarchisation à l'arbre RST qui sera tenue en considération au moment du classement et sélection d'un sous ensemble de phrases pour la génération d'un extrait ciblé pour un utilisateur.
La figure 6.8 montre un exemple d'arbre RST généré suite à une analyse rhétorique. Le résultat est enregistré dans un fichier XML qui englobe les unités minimales (noyau et/ou satellite), les relations rhétoriques qui les relient et le niveau hiérarchique de leurs emplacements dans l'arbre.

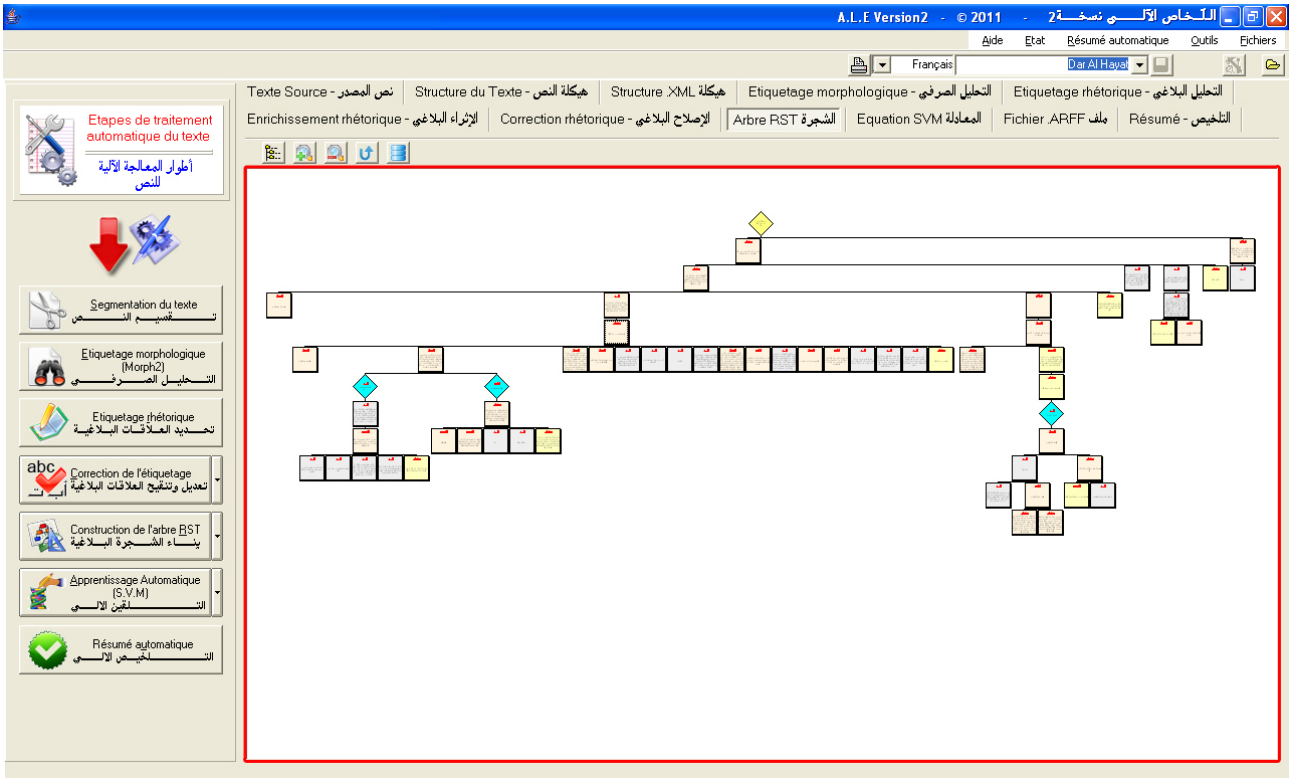

FIGURE 6.8 – Interface de création de l'arbre RST le plus descriptif

6.3.6 Sélection des critères d'extraction reflétant les besoins de l'utilisateur

Notre système de résumé automatique, rappelons-le, se base sur une approche hybride dont les relations rhétoriques guident le processus pour l'identification des informations importantes. Ces relations rhétoriques sont indépendantes d'un domaine particulier.
À ce stade, l'utilisateur intervient pour choisir le type du résumé ou la liste de relations rhétoriques, l'application ou non de la phase d'apprentissage et le taux de réduction à appliquer (en

nombre de phrases).

Le repérage des phrases saillantes à sélectionner dans l'extrait final se base sur des critères d'extraction reflétant les besoins de l'utilisateur (voir figure 6.9) et vont conditionner le futur résumé. Ces critères de sélection sont construits autour des choix ayant des traits numériques et symboliques qui peuvent être perçus grâce aux résultats retournés par les modules cités précédemment. Les relations rhétoriques issues des frames illustrent l'ordre et l'importance des segments textuels. Signalons que les frames rhétoriques sont construits autour des marqueurs linguistiques et des critères morphologiques.

FIGURE 6.9 – Interface utilisateur pour la sélection des critères d'extraction

Ainsi, nous pouvons percevoir que l'extrait final s'oriente vers un extrait dynamique. Cette orientation s'illustre au niveau de la recherche des informations importantes et se fait selon les besoins des utilisateurs.

6.3.7 Sorties du système

À partir des informations sélectionnées et entrées par l'utilisateur, le système "L.A.E - اللّخّاص الآلي" se charge de lancer tous les traitements et modules nécessaires pour afficher l'extrait final.

À ce stade, l'utilisateur visualise le résumé obtenu suite à une intégration des paramètres de génération de l'extrait (e.g. type de résumé, taux de réduction, etc.). Le résumé final présente, dans sa forme globale, le titre principal du texte source, ses sous titres et la liste des phrases saillantes présentées selon un ordre de priorité d'affichage. En revanche, une phrase extraite n'est ainsi, qu'une description partielle de la solution et d'autres phrases auraient dues être extraites en fixant d'autre taux.

La figure 6.10 présente un résumé produit par le système "L.A.E - اللّخّاص الآلي" conformément à la liste des choix donnée au tableau 6.1.

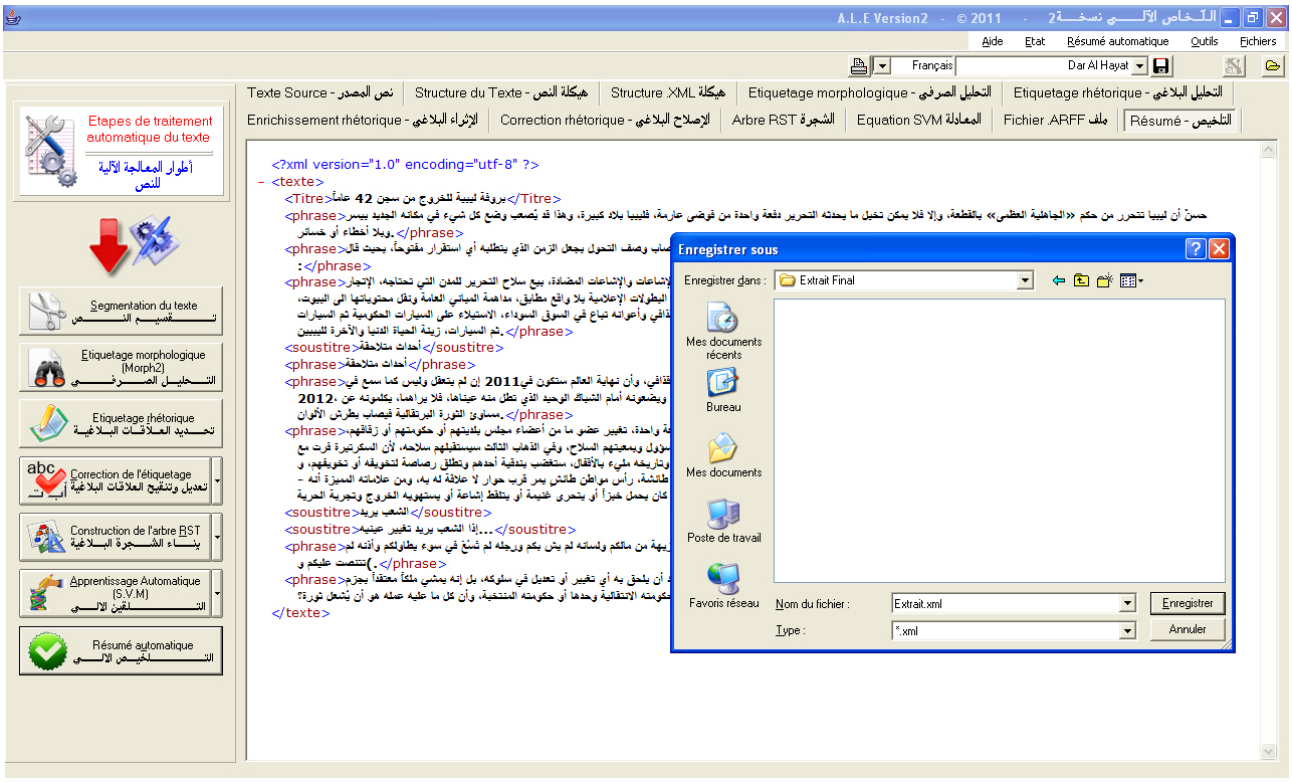

FIGURE 6.10 – Interface de génération de l'extrait final

La figure 6.10 représente l'interface d'enregistrement de l'extrait final. Cet extrait est présenté dans la figure 6.11.

TABLE 6.1 – Critères de choix

Type de critères	Description
Critères symboliques	Type du résumé : indicatif
Critères numériques	Application de l'algorithme SVM
Critères hybrides	Taux de réduction : 10 phrases

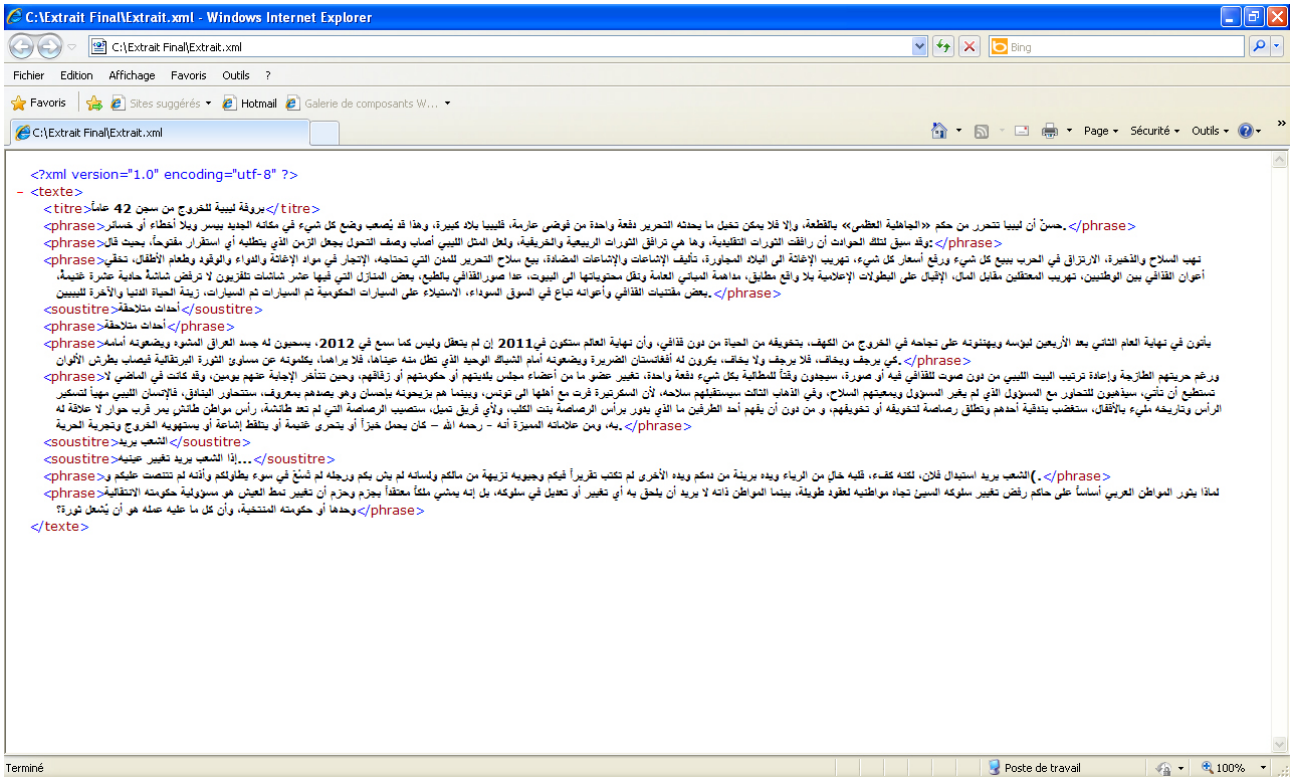

FIGURE 6.11 – Extrait final en format XML

6.4 Conclusion

L'extraction, à partir d'un texte source, des unités textuelles importantes se fait suite à une liste de paramètres choisis par l'utilisateur. Ces paramètres conditionnent le futur extrait généré par le système "L.A.E - اللَخَاص الآلي".

Dans ce chapitre, nous nous sommes intéressés à cette phase. Pour réaliser l'extraction des phrases du résumé, nous avons proposé un système qui se base sur l'approche hybride proposée dans les chapitres 4 et 5.

Les phrases extraites représentent un résumé enregistré dans le fichier au format XML. Signalons qu'un simple changement au niveau des critères et paramètres d'extraction, peut conduire à un

résumé différent. Notre système, permet ainsi la génération d'extraits dynamiques, guidée par le besoin de l'utilisateur.

Dans le chapitre suivant, nous présentons les résultats obtenus par "L.A.E - الـلَخَاص الآلي" suite à l'étude comparative entre les résultats générés par ce système et ceux produits par deux experts humains.

Chapitre 7

Expérimentation et validation

Sommaire

7.1 Introduction 109
7.2 Les métriques d'évaluation utilisées dans les compagnes DUC/TAC . 110
7.2.1 Les mesures ROUGE 112
7.2.2 Les mesures PYRAMID 112
7.2.3 Les mesures de rappel et de précision 113
7.2.3.1 La précision 113
7.2.3.2 Le rappel 114
7.2.3.3 La F-mesure 114
7.2.4 Les mesures d'évaluation linguistique 114
7.3 Évaluation de l'outil d'annotation rhétorique 114
7.3.1 Protocole expérimental 115
7.3.2 Apport de l'étiquetage morphologique dans l'annotation rhétorique 117
7.4 Évaluation du processus de génération d'extrait 118
7.5 Conclusion 119

7.1 Introduction

L'évaluation est une étape essentielle dans le développement d'une application informatique pour le TALN, et en particulier les systèmes de résumé automatique. Elle concerne le contenu et la qualité des résumés produits, afin d'estimer la capacité d'une application à effectuer des tâches qu'on lui soumet.

L'évaluation de la qualité des résumés produits teste la lisibilité, la grammaticalité et la cohérence du résumé. Toutefois, ces évaluations peuvent être faites d'une façon *intrinsèque* ou *extrinsèque* [Jones 1996].

Une évaluation *intrinsèque* mesure la qualité du résumé machine en lui-même d'après ses propriétés et son contenu. Ceci est fait en mesurant la couverture des idées clés du document source dans le résumé ou en comparant le contenu du résumé automatique avec un résumé

idéal produit par un expert [Cole 1996].
Une évaluation *extrinsèque* permet de connaître si le résumé permet de répondre aux attentes des utilisateurs dans certaines tâches. Les utilisateurs doivent, par exemple, décider si les résumés permettent ou non de déterminer les thèmes abordés par les textes sources. Bien souvent, les utilisateurs, peuvent en fonction d'une requête classique de recherche liée à un thème particulier, évaluer la pertinence du document grâce au résumé. L'efficacité d'un résumé est jugée par sa couverture maximale, de tout le document et par son aide à répondre à toutes les questions portant dessus, comme cela serait possible avec la lecture du texte original. Les expériences de [Jing 1998] ont montré comment différents paramètres tels que la longueur du résumé peuvent affecter le résultat de l'évaluation.
Certains types d'évaluation peuvent être abordés en utilisant les mesures de rappel et de précision. D'autres doivent être accomplis par des juges qualifiés. Dans le cadre de cette thèse, l'objectif de notre évaluation est de comparer des résumés produits automatiquement par notre système avec des extraits construits, par deux experts humains, à partir des mêmes textes. L'évaluation s'effectuera ainsi, par des mesures déterminées par les paramètres rappel et précision.
Dans ce chapitre, nous allons d'abord donner un bref aperçu sur les métriques d'évaluation les plus utilisées dans le domaine du résumé automatique. Puis, nous présentons les résultats d'évaluation de l'outil d'annotation rhétorique dans la tâche d'assignation des relations rhétoriques. Les performances du module d'annotation ont été testées pour cent articles du corpus. Par la suite, nous présentons l'évaluation de notre système qui se base sur une étude comparative mettant en jeu les résumés du système avec les résumés humains. Cette étude comparative consiste à identifier les phrases communes entre les résultats générés par notre système avec ceux réalisés par deux experts humains. Signalons que notre évaluation porte sur cinquante textes, choisis au hasard, pris du corpus test constitué d'articles de presse provenant du site Dar el Hayat.

7.2 Les métriques d'évaluation utilisées dans les compagnes DUC/TAC

Depuis 2001, le NSIT [1] "National Institute for Science and Technology" organise la compagne d'évaluation DUC [2] "Document Understanding Conference". La conférence DUC a pour objectif de promouvoir les recherches dans le domaine de résumé et d'extraction automatiques. En 2008, la conférence DUC a rejoint les conférence TREC "Text Retrieval Conferences" pour

1. http://www-nlpir.nist.gov/projects/duc
2. http://duc.nist.gov

former une seule conférence TAC[3] "Text Analysis Conference". Dans cette même année, la compagne TAC a organisé un atelier portant sur trois volets : *Summarization*, *Question & Answering* et *Recognizing Textual Entailment*.
Plusieurs sessions DUC et TAC se sont tenues jusqu'à présent. Nous présentons dans le tableau 7.1 les tâches introduites par ces compagnes afin de promouvoir les progrès réalisés dans le domaine du résumé automatique de documents.

TABLEAU 7.1 – Les tâches d'évaluation proposées dans les conférences DUC et TAC [Torres-Moreno 2011]

Conférence	Tâche de résumé
DUC 2001	Résumés génériques mono et multi-documents
DUC 2002	Résumés génériques mono et multi-documents
DUC 2003	Résumés Courts (*headlines*) et résumé multi-documents
DUC 2004	Résumés Courts multilingues multi-documents et biographiques
DUC 2005	Résumés orientés
DUC 2006	Résumés orientés multi-documents
DUC 2007	Résumés orientés avec un besoin utilisateur, multi-documents, d'une taille inférieure ou égale à 250 mots, à partir d'un regroupement d'environ 25 documents ; résumés mise à jour (*Pilot Task*)
TAC 2008	Résumés mise à jour (*Update task summarization*) orientés par un besoin utilisateur et multi-documents
TAC 2009	Résumés mise à jour (*Update task summarization*) orientés par un besoin utilisateur et multi-documents
TAC 2010	Résumés orientés multi-documents (*Guided summarization*) et la tâche *Automatically Evaluating Summaries Of Peers*
TAC 2011	Résumés orientés multi-documents (*Guided summarization*), *Automatically Evaluating Summaries Of Peers*, et la nouvelle tâche *MultiLing pilot* pour favoriser et promouvoir l'utilisation d'algorithmes multilingues pour le résumé. Cela comprend l'effort de transformation d'un algorithme ou d'un ensemble de ressources monolingues en une version multilingue.

Dans ce qui suit, nous présentons les principales mesures adoptées par la communauté lors de l'évaluation des systèmes de résumé automatiques.

3. http://www.nist.gov/tac

7.2.1 Les mesures ROUGE

Les mesures produites par ROUGE "Recall-Oriented Understudy for Gisting Evaluation" [4] sont des mesures automatiques dont le calcul s'appuie sur la comparaison du résumé système avec plusieurs résumés de référence. Cette comparaison se base sur la co-occurence des ngrammes.
La formule, proposée pour les métriques générées par ROUGE, utilise une moyenne pondérée des ngrammes à longueurs variables et extraits à partir des résumés systèmes et un ensemble de résumés de référence. La formule générale de ces métriques est la suivante :

$$ROUGE\_n = \frac{\sum_{S \in R_{ref}} \sum_{n\_gammes \in S} Co\_occurrences(R_{can}, R_{ref})}{\sum_{S \in R_{ref}} \sum_{n\_gammes \in S} nombre(n\_gammes)} \tag{7.1}$$

où Co_occurrences(R_{can},R_{ref}) correspond au nombre maximum de Co_occurrences de n_gammes dans un résumé candidat R_{can} et l'ensemble de résumés de référence R_{ref} et nombre (n_gammes) est la somme totale du nombre de n_grammes présents dans les résumés de référence R_{ref}. Notons que des études d'évaluation ont montré que la mesure ROUGE avec des bi-grammes où $n = 2$ et ROUGE_SU4 [5] "Rouge with Skip Unit" sont parmi les métriques ROUGE qui ont une meilleure corrélation avec les jugements humains [Lin 2004].
Afin de tenir compte des informations syntaxiques, d'autres mesures ont été développées. Ces mesures utilisent des informations syntaxiques permettant d'identifier les relations entre les bi-grammes (e.g. BE [6]" Basic Elements" [Hovy 2006], PYRAMID [Nenkova 2007]).

7.2.2 Les mesures PYRAMID

Les mesures générées par PYRAMID se basent sur le découpage des résumés de référence en unités d'informations sémantiques SCU "Semantic Content Unit" élémentaires, puis sur la vérification de la présence de ces unités dans le résumé système [Nenkova 2007]. L'idée consiste à déterminer, à partir des résumés de référence, des unités sémantiques SCU [7] et qui expriment, de diverses manières, une notion unique, un poids qui dépend de sa présence dans chacun des résumés de référence. Ainsi, on construit une pyramide dont le haut est occupé par les unités communes entre les résumés de référence. Il s'agit ensuite de déterminer, dans le résumé système à évaluer, les unités qui correspondent à celles dans la pyramide.

4. http://haydn.isi.edu/ROUGE

5. ROUGE_SU4 autorise la présence d'une distance inférieur à 4 mots séparant les deux mots d'un bi-gramme.

6. Les Basic Elements se constituent des triplets (H | M | R) où H = Tête (*Head*), M = Modifieur (*Modifier*) et R = Relation (*Relation*) qui lie les H avec les M [Hovy 2005].

7. Une SCU se compose d'un ensemble d'éléments qui, dans leur contexte de phrase, expriment le même contenu sémantique [Nenkova 2004].

La somme normalisée des scores des unités est présentée par l'équation suivante :

$$PYRAMID = \frac{\sum poids(SCU_{candidat})}{\sum poids(SCU_{rfrences})} \tag{7.2}$$

Notons que PYRAMID est une méthode semi-automatique d'évaluation des résumés développée par NenKova et al. [Nenkova 2004] et ayant pour objectif de surmonter le problème de sémantique non abordé par ROUGE.

7.2.3 Les mesures de rappel et de précision

Le rappel et la précision présentent des mesures de similarité classiques de recherche d'information. Ces mesures issus de cette discipline, ont pour objectif d'indiquer à quel point un système obtient des performances proches à celles obtenues manuellement par des humains [Torres-Moreno 2011].
Rappelons que ces deux mesures sont calculées, pour une requête qui cherche des phrases dans un fond textuel, à partir des trois paramètres suivants :

- P : nombre de phrases non pertinentes fournies par le système,
- Q : nombre de phrases pertinentes fournies par le système,
- R : nombre de phrases pertinentes présentes dans un fond textuel et non fournies par le système.

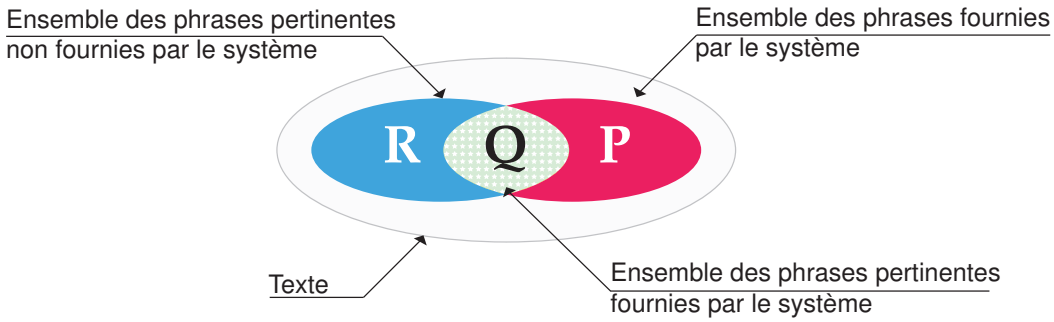

FIGURE 7.1 – Paramètres de calcul de la précision et du rappel [Minel 2002a]

Dans ce qui suit, nous donnons les formules permettant de mesurer respectivement le rappel, la précision et la F-mesure.

7.2.3.1 La précision

La précision consiste à évaluer le niveau du bruit du système. Le bruit est la proportion complémentaire, celle de mauvaises réponses [Saggion 2000]. Formellement, la précision est le quotient suivant :

$$Précision = \frac{Q}{(Q+P)} \tag{7.3}$$

7.2.3.2 Le rappel

Le rappel permet de mesurer le nombre de phrases correctes détectées par rapport au nombre total des phrases type recherchées dans l'ensemble du texte. Le silence (rappel) est le complémentaire, la proportion des phrases oubliées par le système [Torres-Moreno 2011].

$$Rappel = \frac{Q}{(Q+R)} \tag{7.4}$$

7.2.3.3 La F-mesure

Afin de pondérer l'importance de chacun de ces deux paramètres (la précision et le rappel), un troisième paramètre, appelé la F-mesure, est en général calculé à partir de ces deux paramètres [Minel 2002a].

$$F - mesure = \frac{2 * Précision * Rappel}{(Précision + Rappel)} \tag{7.5}$$

7.2.4 Les mesures d'évaluation linguistique

Les mesures d'évaluation linguistique [8] sont des évaluations manuelles et ont pour objectif de mesurer la grammaticalité, la redondance, les ruptures de cohérence et la clarté référentielle. Ces mesures sont attribuées suite à des juges humains qui, après lecture de l'extrait candidat, donnent une note de 1 à 5 pour chaque critère : 1 très mauvais (*Very Poor*), 2 pauvre (*Poor*), 3 suffisant (*Barely Acceptable*), 4 bon (*Good*) et 5 très bon (*Very Good*).

7.3 Évaluation de l'outil d'annotation rhétorique

Le développement de notre outil d'annotation rhétorique a subi deux versions. Dans une première version [Maaloul 2010a] de l'outil, la détection des relations rhétoriques repose exclusivement sur des frames rhétoriques basés sur des indices linguistiques.
La nouvelle version de cet outil [Maaloul 2012a], est plus "approfondie" que la première. En ce sens, l'outil est devenu capable de résoudre certains cas d'ambiguïtés liés à l'absence d'indices

8. http://duc.nist.gov/duc2005/quality-questions.txt

complémentaires au voisinage de l'indicateur déclencheur, et qui sont utiles à la confirmation du concept énoncé par l'indicateur déclencheur. Ainsi, l'amélioration se résume dans l'utilisation des critères morphologiques qui ont montré leurs utilités dans le repérage des relations rhétoriques. D'autres, mis-à-jour, de moindre importance, mais tout aussi utiles, ont concerné l'ajout de fonctionnalités dans le but d'assigner un arbre RST personnalisé [9] et par conséquent convenir au mieux aux besoins de l'utilisateur au moment de la sélection des phrases formant le contenu de l'extrait généré.

Dans cette section, nous allons évaluer les performances des frames rhétoriques basés sur des critères morphologiques dans la tâche d'assignation des relations rhétoriques. Les performances de l'outil d'annotation ont été testées pour cent articles de presse, pris au hasard, du corpus de test. La valeur relativement réduite du corpus de test s'explique par le coût élevé de l'analyse discursive, de qualité, des articles en terme de temps vu qu'elle se base sur une compréhension approfondie du contenu et du domaine traité dans le texte. Ainsi, moins le lecteur, l'expert dans notre cas, a des connaissances sur le domaine traité moins son exigence vis-à-vis de la lisibilité est forte, ce qui reflète par ailleurs la qualité, l'importance et le nombre de relations rhétoriques détectées.

Pour remédier au maximum à ce problème, nous avons choisi d'évaluer cet outil sur la base des jugements réalisés par deux linguistes. Ces derniers, ont annoté le corpus test en découpant les phrases, des articles, en segments et en associant, par la suite, à ces segments les relations qui les relient. De même, et pour résoudre le problème de désaccord entre les jugements, nous avons utilisé l'indice statistique Kappa [Cohen 1960] afin de mesurer le degré d'accord inter-annotateur.

7.3.1 Protocole expérimental

Pour bien apprécier les performances de notre outil d'annotation rhétorique dans sa tâche de détection des relations rhétoriques, nous avons choisi de tester les résultats générés automatiquement avec ceux déterminés suite à des jugements donnés par deux experts.

Une mesure du degré d'accord est ainsi nécessaire pour déterminer la conformité de ces jugements donnés par les experts.

Le taux d'accord ou de "*concordance*" est estimé par le coefficient K Kappa défini par Cohen [Cohen 1960]. Ainsi, l'accord observé entre deux jugements est présenté par le coefficient Kappa K qui est le pourcentage d'accord maximum corrigé de ce qu'il serait sous le simple effet du hasard. C'est un nombre réel, sans dimension, compris entre {-1} et {1} [Carletta 1996].

Le calcul d'accord entre deux observateurs statistiquement indépendants se présente comme

9. L'utilisateur peut modifier la position d'un item de l'arbre RST, élaborer une nouvelle relation de ce dernier avec un autre, etc.

suit :

$$K = \frac{P_0 - P_e}{1 - P_e} \tag{7.6}$$

avec :

- P_0 : Le taux de concordance d'accord dans les réponses communes données par les deux experts.
- P_e : Le taux d'accord aléatoire ou concordance aléatoire, calculé par la somme des produits des normes de chaque classe de chaque expert, divisée par le carré du nombre total de réponses à ramener.

$$P_e = \frac{1}{N^2} \sum_{i=1}^{r} (n_i * n_i) \tag{7.7}$$

où :

- r représente le nombre de modalités,
- n est le vecteur des proportions observées d'un tableau de contingence à r modalités, et
- N est le nombre total de réponses à ramener.

Selon les expériences de Landis et Koch [Landis 1977], ils ont pu déduire un échelle de classement pour l'indice kappa (K) afin de juger le degré d'accord (voir tableau 7.2).

TABLEAU 7.2 – Degré d'accord et valeur de Kappa proposés par Landis et Koch [Landis 1977]

Accord	**Kappa**
Excellent accord	0,81 $\leq K \leq$ 1,00
Bon accord	0,61 $\leq K \leq$ 0,80
Accord modéré	0,41 $\leq K \leq$ 0,60
Accord médiocre	0,21 $\leq K \leq$ 0,40
Mauvais accord	0,00 $\leq K \leq$ 0,20
Très mauvais accord	-1 $\leq K \leq$ 0,00

Dans notre expérimentation, l'indice Kappa (K) moyen[10] est de 0,78 %, ce qui correspond à un bon accord d'après l'échelle de Landis et Koch.

10. L'indice Kappa moyen est déterminé à partir de la moyenne des indices Kappa obtenus par l'application de l'équation 7.6 sur les cent articles de presse du corpus test

7.3.2 Apport de l'étiquetage morphologique dans l'annotation rhétorique

Afin d'évaluer l'importance de l'annotation rhétorique proposée dans le chapitre 4, nous avons procédé à l'évaluation de la performance et la pertinence de cette dernière à l'aide d'une étude comparative qui met en jeu les résultats générés par notre système avec ceux des deux experts. Signalons que cette évaluation est réalisée par la première version (V_1) de notre outil d'annotation [Maaloul 2010a] qui repose uniquement sur des frames rhétoriques basés sur des indices linguistiques.

Le tableau 7.3 des résultats des performances de l'outil d'annotation rhétorique, permet de présenter le nombre de relations rhétoriques trouvées par les experts humains (E), le nombre total de relations rhétoriques trouvées par l'outil d'annotation version 1 (D), le nombre de relations rhétoriques correctement trouvées par l'outil d'annotation version version 1 (C). De même, pour la deuxième version (V_2) de l'outil d'annotation qui repose sur des frames rhétoriques basés sur des indices linguistiques et des critères morphologiques (voir paragraphe 4.3 du chapitre 4).

Le tableau 7.3 se lit ligne par ligne comme suit :

Sur cent documents comportant 2083 relations rhétoriques déduites par les experts, l'outil d'annotation version (V_1) a détecté 945 relations parmi lesquelles 646 relations sont correctement déduites, 299 relations ont été déduites en tant que relations différentes à celles présentées par les experts, se qui donne un pourcentage de performance de 31.01%, etc.

Tableau 7.3 – Résultats des performances de l'outil d'annotation rhétorique

Experts humains	**Outil d'annotation V_1**		**Outil d'annotation V_2**	
(E)	(D)	(C)	(D)	(C)
2083	945	646	1181	811
Pourcentage de performance	31.01%		38.93%	

Les résultats de l'évaluation de l'outil d'annotation, montrent clairement l'apport des frames rhétoriques basés sur des critères morphologiques dans la tâche d'annotation des relations rhétoriques. Nous remarquons ainsi, que le pourcentages des relations correctement déduites s'élève à un taux de 38.93% avec une augmentation de 7,92% par rapport à la première version qui utilise uniquement des frames rhétoriques basés sur des indices linguistiques.

Toutefois, nous pouvons remarquer que l'utilisation des frames rhétoriques basés sur des critères morphologiques augmente le nombre de relations détectées par 236 relations, dont 165 relations correctement déduites, ce qui implique que 69,91% des nouvelles relations déduites sont correctes.

7.4 Évaluation du processus de génération d'extrait

L'évaluation des systèmes de résumé constitue un sujet émergent. Elle concerne le contenu et la qualité des résumés produits. L'évaluation du contenu vérifie si le système automatique est capable d'identifier les thèmes principaux du document source. Alors que celle de la qualité teste la lisibilité et la cohérence du résumé [Minel 2002a].

Dans le cadre de ce travail, nous nous intéressons à l'évaluation du contenu des résumés. Nous ne nous intéressons pas à l'évaluation de la qualité vu que la méthode que nous avons proposée tout au long de ce travail s'intéresse principalement à la génération d'extraits et non pas de résumés. Ainsi, l'évaluation du contenu des résumés générés par notre système "L.A.E - الـلَخَاص الآلي", est réalisée à l'aide d'une étude comparative mettant en jeu les résultats générés par notre système avec ceux réalisés par deux experts humains.

Cette étude fait appel au calcul des métriques standards d'évaluations à savoir, le rappel et la précision et la F-mesure vu que nous ne disposons pas de résumés de référence mais d'extraits de référence (pour chaque article de presse, les experts ont sélectionné des phrases qui selon eux peuvent être incluses dans le résumé).

Le corpus de test utilisé est constitué de 150 articles de presse rapatriés du site web du journal quotidien généraliste arabophone "Dar al Hayat - دَار الحيَاة" [11]. Ce corpus est différent des deux corpus d'étude et d'apprentissage présentés précédemment dans le chapitre 4 et 5.

À chaque article du corpus correspond un résumé de référence produit manuellement par deux experts humains. Un résumé de référence est un ensemble de phrases pertinentes extraites par les experts, à partir des articles sources.

Une phrase est considérée comme pertinente si elle a été extraite par les deux experts (accord total entre les deux experts) ou si elle obtient un indice de Kappa supérieure à 0.68%.

Afin de montrer l'efficacité de l'approche hybride proposée pour le résumé automatique, nous avons procédé à l'évaluation du système "L.A.E - الـلَخَاص الآلي" de deux manières.

La première évaluation est réalisée par L.A.E (V_0) basé uniquement sur une approche symbolique (la sélection des phrases de l'extrait final se base uniquement par l'analyse RST).

La deuxième évaluation est réalisée par L.A.E (V_1) sur une approche hybride (la sélection des phrases de l'extrait final est basée sur l'analyse RST et l'apprentissage).

Le tableau 7.4 montre que L.A.E (V_1) est plus performant que L.A.E (V_0). En effet, le système L.A.E (V_1) a obtenu une F-mesure de 0.53 alors que le système L.A.E (V_0) a obtenu uniquement 0.21 pour cette même mesure.

Ces résultats confirment la capacité de l'approche hybride à améliorer les performances du système de résumé automatique.

Ainsi, l'idée qui postule que le « résumé idéal » nécessite principalement une analyse linguistique

11. http://www.daralhayat.com

TABLEAU 7.4 – Moyennes des mesures de Rappel, Précision et F-mesure

Systèmes	Rappel	Précision	F-mesure
"L.A.E - الـلَخَاص الآلي" V_0"	0,21	0,32	0,25
"L.A.E - الـلَخَاص الآلي" V_1"	0,47	0,61	0,53

inhérente aux méthodes basées sur la compréhension du texte, n'a pas été approuvée compte tenu de ce qui est obtenu précédemment. Par ailleurs, l'élimination de l'aspect numérique réduit considérablement la précision du système "L.A.E - الـلَخَاص الآلي" lors de la phase de sélection des phrases formant l'extrait final.
Dans le même ordre d'idées, l'analyse rhétorique doit opérer aussi davantage à d'autres phases d'enrichissement des relations rhétoriques. Notons que le taux de performance de l'outil d'annotation rhétorique a prouvé seulement un taux de performance de 38,93% (voir tableau 7.3).
Enfin, l'utilisation du mécanisme d'apprentissage a permis de remédier au problème de manque d'information linguistique. En effet, rappelons-le, lorsque l'analyse rhétorique n'est pas capable de déterminer la nature d'une relation (cas d'absence d'indice déclencheur), l'apprentissage intervient pour résoudre le problème et détermine si la phrase concernée par cette relation ("Autres - آخر") est pertinente ou non.
Les résultats auxquels nous avons abouti contribuent bien à l'orientation suivie dans l'approche hybride proposée. Toutefois, l'étape de sélection des phrases pertinentes, par le biais d'apprentissage, nécessite la présence d'un corpus de test constitué par des articles de presse munis de leurs résumés de références.
L'absence d'un tel corpus dans le mécanisme d'apprentissage inflige au concepteur la lourde tâche de classer les phrases saillantes non porteuses d'information linguistique.

7.5 Conclusion

Dans ce chapitre, nous avons décrit, dans un premier temps, quelques mesures d'évaluation des résumés tenues par la communauté scientifique afin de donner une idée sur la qualité d'un résumé candidat.
Rappelons, qu'un des points faibles commun entre les mesures d'évaluation étant la nécessité d'un ou de plusieurs résumés de référence, chose qui est généralement coûteuse en temps et en ressources.
Dans un deuxième temps, nous avons présenté une illustration des résultats obtenus, suite à

une évaluation, de l'outil d'annotation rhétorique. Les résultats de cette évaluation nous ont permis de corriger et d'ajouter des frames rhétoriques basés sur des critères morphologiques. Cette amélioration a été approuvée par les résultats de performance de l'outil d'annotation rhétorique (V_2) basé sur des frames rhétoriques utilisant des indices linguistiques et des critères morphologiques (voir tableau 7.3).

Dans un troisième temps, nous avons présenté les résultats de l'évaluation du système "L.A.E - الـلَخَاص الآلي".

Notons que la grande difficulté à laquelle est confrontée la tâche d'évaluation réside dans l'absence de la notion de résumés modèles ou de standards ; ce qui a nécessité la production d'un corpus de résumé de référence par deux experts.

Les résultats obtenus en faveur du système "L.A.E - الـلَخَاص الآلي" V_1 (une version du système basée sur l'approche hybride) sont encourageants et prouvent la performance de l'approche hybride proposée. Ces résultats, ont montré, en premier lieu, l'applicabilité de l'approche dans le contexte de mono-document sans restriction quant à leur thème (Éducation, Sport, Science, Politique, Reportage, etc.), leur contenu et leur volume. Ils ont aussi montré l'importance de l'apprentissage dans la phase de classement et sélection des phrases formant l'extrait final.

L'étape d'apprentissage que nous avons intégrée a permis, aussi, de définir une nouvelle manière d'agréger des critères permettant de qualifier l'importance de l'extrait.

Conclusion et perspectives

La problématique du résumé automatique qui a été abordée dans cette thèse s'est cristallisée autour de deux points. Le premier point concerne les critères utilisés pour décider du contenu essentiel à extraire. Le deuxième point se focalise sur les moyens qui permettent d'exprimer le contenu essentiel extrait sous la forme d'un texte ciblant les besoins potentiels d'un utilisateur. Pour comprendre comment les résumés automatiques sont produits, nous avons présenté, dans le premier et le deuxième chapitres, une étude sur les méthodes existantes pour résoudre ce problème. Nous avons constaté trois approches ; l'approche numérique qui utilise des méthodes de calcul, simplement implantées et rapidement adaptables à d'autres domaines, mais limitée dans le sens où elle n'a pas une vision globale du texte ; l'approche symbolique fondée sur des connaissances avancées et qui ne peut être appliquée qu'à des domaines restreints et l'approche hybride qui tient compte des traits du discours et qui est plus prometteuse. Malgré les avancées enregistrées grâce à cette dernière approche, le champ d'investigation reste ouvert.

Afin de comprendre la particularité de la langue arabe, qui est considérée comme une langue difficile à maîtriser dans le domaine du TALN, nous avons abordé, dans le troisième chapitre, l'étude des méthodes et approches existantes pour résoudre les problèmes de segmentation, d'étiquetage grammatical, d'analyses morphologique et syntaxique.

Signalons que les résumés produits par les méthodes d'extraction sont constitués par des phrases généralement déconnectées du besoin des utilisateurs. Nous avons, alors, visé la génération de résumés automatiques en se basant sur une double clause ; premièrement à travers une vision globale du texte source et deuxièmement en tenant compte des besoins potentiels d'un utilisateur. Nous avons décidé de baser l'extraction du contenu essentiel sur des recommandations fournies par le type du résumé choisi par l'utilisateur (type informatif, indicatif, opinion, etc.). Par ailleurs, nous avons abordé les questions de la structuration rhétorique des textes en se basant sur la théorie de la structure rhétorique – RST, qui fait appel à des frames rhétoriques basés sur des indices linguistiques et des critères morphologiques. Ces frames ont un double rôle. D'abord, ils lient deux unités minimales adjacentes, dont l'une possède le statut de noyau et l'autre le statut de noyau ou de satellite, et ensuite ils déterminent le type de relations qui les relient.

L'étude du corpus de travail nous a mené vers la définition de ces frames rhétoriques. Elle nous a permis aussi de mettre en place des relations rhétoriques dans lesquelles les frames sont organisés.

De plus, nous avons proposé une approche hybride basée sur l'utilisation conjointe d'une analyse symbolique et d'un traitement numérique.

L'analyse symbolique consiste en une analyse rhétorique (voir chapitre 4) qui a pour rôle de

mettre en évidence, dans un texte source, les unités minimales (noyaux – segments de texte primordiaux pour la cohérence et qui charpentent un discours) nécessaires pour le processus d'extraction. Cela est effectué dans une optique d'utiliser un traitement numérique basé sur l'algorithme d'apprentissage SVM.

Cet apprentissage permet de déterminer parmi les phrases non porteuses d'information rhétorique et ayant la relation "Autres - آخر" (i.e. la relation rhétorique "Autres - آخر" est attribuée lorsqu'aucune relation rhétorique n'est déterminée) celles qui sont pertinentes pour l'extrait final.

L'approche hybride, que nous avons proposée pour la génération d'extraits, se compose de quatre étapes à savoir : la segmentation du texte source en plusieurs unités textuelles plus petites, l'étiquetage morphologique des unités textuelles segmentées, l'analyse rhétorique, et la sélection et le classement des unités textuelles pertinentes selon un profil utilisateur.

Afin de montrer la faisabilité de notre approche, nous avons développé le système "L.A.E (**L**axAS **A**l **E**l|ly) - الـلَخَاص الآلي". Notre implantation se base, en partie, sur l'utilisation d'un certain nombre d'outils (le segmenteur **STAr** [Belguith 2005] et l'analyseur morphologique **Morph-2** [Chaaben 2010]).

Pour représenter les types d'informations auxquels nous nous intéressons, les frames associés aux relations rhétoriques ont été transposés en règles d'exploration contextuelle que nous utilisons pour les tâches de détection des unités noyaux et des relations rhétoriques. Ces tâches sont utilisées pour hiérarchiser le texte sous forme d'un arbre RST, en vue de le réduire au niveau de l'étape de sélection et de classement des unités textuelles. Cette réduction se fait en fonction de différents critères (e.g. type de résumé, taux de réduction, etc.) définis selon un besoin utilisateur.

Nous avons évalué l'approche proposée sur un corpus de test constitué de 150 articles de presse rapatriés du site web du journal quotidien généraliste arabophone "Dar al Hayat - دَار الحيَاة". Rappelons qu'à chaque article du corpus correspond un résumé de référence [12] produit manuellement par deux experts humains.

L'évaluation a donné les valeurs de 0.47, 0.61 et 0.53 respectivement pour les mesures de rappel, précision et F-mesure.

Un des points forts de l'approche proposée réside dans deux aspects :

- L'interaction de connaissances linguistiques avec d'autres numériques,
- Le traitement d'articles de presse portant sur des thèmes hétérogènes.

Il est clair que l'utilisation du mécanisme d'apprentissage a permis de remédier au problème de manque d'information linguistique. Il a ainsi réduit le problème d'une phrase pertinente non retenue dans l'extrait final car non porteuse de relation rhétorique.

12. Un résumé de référence est un ensemble de phrases pertinentes extraites par les experts, à partir des articles sources.

Cependant, ce problème de perte d'information rhétorique apparaissant de façon récurrente lors de l'analyse rhétorique à cause de la non-similarité surfacique aux frames rhétoriques, nécessite la reconnaissance d'autres frames rhétoriques pour améliorer les résultats. Or, la tâche de trouver manuellement des marqueurs linguistiques est coûteuse.

Dans nos travaux futurs, nous souhaitons explorer l'utilisation de méthodes d'acquisition automatique de connaissances et de méthodes de classification statistique afin de surmonter cette faiblesse actuelle au niveau de la détection des relations rhétoriques.

Un autre problème à résoudre est la présence de phrases similaires dans un même extrait final, c'est-à-dire qui ont une même sémantique. La nécessité d'introduire une technique de détection des phrases redondantes dans un résumé émerge alors.

Enfin, nous concluons en annonçant qu'un thème de recherche relatif à la production de résumé mono-document dans plusieurs langues et qui traite l'apport des méthodes symboliques dans la génération automatique d'extrait est en train d'être exploré en linguistique informatique. Ceci nous incite à explorer la possibilité d'utiliser d'autres méthodes symboliques, jugées plus performants, dans l'analyse linguistique et qui assument une compréhension profonde du document afin d'assurer par la suite la génération d'un résumé de qualité, présentant le contenu du texte, les intentions de l'auteur et les besoins de l'utilisateur.

Liste des publications de la thèse

Chapitre d'un livre

1. "A paraître en 2013 [Belguith 2013]"
   Title : Automatic summarization of Semitic languages
   Authors : Lamia Hadrich Belguith, Mariem Ellouze, Mohamed Hédi Maaloul, Maher Jaoua, Fatma Kallel Jaoua and Philippe Blache
   Abstract : This chapter addresses automatic summarization of Semitic languages. After a presentation of the theoretical background and current challenges of the automatic summarization, we present different approaches suggested to cope with these challenges. These approaches fall on to two classes : single vs. multiple document summarization approaches. The main approaches dealing with Semitic languages (mainly Arabic, Hebrew, Maltese and Amharic) are then discussed. Finally, a case study of a specific Arabic automatic summarization system is presented. The summary section draws the most insightful conclusions and discusses some future research directions.

Articles publiés dans des conférences internationales

1. [Maaloul 2012b] : Mohamed Hédi Maaloul, Iskandar Keskes, Mohamed Mahdi Boudabous and Lamia Hadrich Belguith, *Étude comparative entre trois approches de résumé automatique de documents arabes.* Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2 : TALN, pp 225–238, Grenoble, France, 2012.
2. [Maaloul 2012a] : Mohamed Hédi Maaloul, Wajdi Ajjel et Lamia Hadrich Belguith, *Apport de l'analyse morphologique dans la détection des relations rhétoriques* - دور التحليل اللُغوي في رصد العلاَقات البلَاغية., International Conference on Arabic Language Processing' 2012 (CITALA'2012), Rabat, Morocco, 2012.
3. [Maaloul 2011] : Mohamed Hédi Maaloul, Keskes Iskander, Boudabous Mohamed Mahdi and Hadrich Belguith Lamia, *Le résumé automatique de documents arabes : entre une technique d'apprentissage et la théorie de la structure rhétorique* - التلخيص الآلي لِلنصوص العربية : بين نظرية التلقين الرقمي و نظرية البنية البلَاغية., 7th International Computing Conference in Arabic 2011 (ICCA'2011), Riyadh, Saudi Arabia, 2011.
4. [Maaloul 2010a] : Mohamed Hédi Maaloul et Iskandar keskes, *Résumé automatique de documents arabes basé sur la technique RST.* 12èmes Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (JEP-TALN-RECITAL'2010), Québec, Canada, 2010.

5. [Maaloul 2010c] : Mohamed Hédi Maaloul and Iskander Keskes and Lamia Hadrich Belguith and Philippe Blache, *Automatic summarization of arabic texts based on RST technique*. 12th International Conference on Enterprise Information Systems (ICEIS'2010), Funchal, Madeira - Portugal, 2010.
6. [Maaloul 2010b] : Mohamed Hédi Maaloul, Iskander keskes and Lamia Hadrich Belguith, *Résumé automatique de documents arabes basé sur la théorie de la structure rhétorique* - التلخيص الآلي لِلنصوص العربية اعتمَادًا علَى نظرية البنية البلَاغية., 6th International Computing Conference in Arabic 2010 (ICCA'2010), Hammamet, Tunisia, 2010.
"Cet article a obtenu le prix du meilleur article et de la meilleure présentation."
7. [Maaloul 2010d] : Mohamed Hédi Maaloul, Mohamed Mahdi Boudabous and Lamia Hadrich Belguith, *Digital learning for summarizing Arabic documents*. 7th international conference on Advances in natural language processing IceTAL'2010, pp 79–84, Publisher Springer-Verlag, Berlin, Heidelberg, 2010.

Bibliographie

[Abraham 1992] Maryvonne Abraham et Jean-Pierre Desclès. *Interaction between lexicon and image : linguistic specifications of animation.* In Proceedings of the 14th conference on Computational linguistics - Volume 3, COLING '92, pages 1043–1047, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. (Cité en page 40.)

[Aliod 1998] Diego Mollà Aliod, Jawad Berri et Michael Hess. *Extraction automatique de réponses : implémentation du système ExtrAns.* In Proceedings of the fifth conference TALN 1998 (Traitement Automatique des Langues Naturelles), pages 12–21, Paris, France, JUN 1998. (Cité en page 40.)

[Alrahabi 2006a] Motasem Alrahabi, Brahim Djioua et Jean-Pierre Desclés. *Annotation Sémantique des Énonciations en Arabe.* INFORSID'2006 - Hammamet –Tunisie, 2006. (Cité en pages 1, 66 et 68.)

[Alrahabi 2006b] Motasem Alrahabi, Amr Helmy Ibrahim et Jean-Pierre Desclés. *Semantic Annotation of Reported Information in Arabic.* In FLAIRS Conference'06, pages 263–268, 2006. (Cité en page 2.)

[Alrahabi 2008] Motasem Alrahabi et Jean-Pierre Desclés. *Automatic Annotation of Direct Reported Speech in Arabic and French, According to a Semantic Map of Enunciative Modalities.* In Bengt Nordström et Aarne Ranta, editeurs, Advances in Natural Language Processing, volume 5221 of *Lecture Notes in Computer Science*, pages 40–51. Springer Berlin / Heidelberg, 2008. (Cité en page 66.)

[Alrahabi 2009] Motasem Alrahabi et Jean-Pierre Desclés. *Opérations de prise en charge énonciative : assertion, médiatif et modalités dans le discours rapporté direct, en arabe et en français.* In in : Methods of lexical analysis, theoretical assumptions and practical applications, 2009. (Cité en page 66.)

[Alrahabi 2010] Almoatasem Alrahabi. Excom-2 : plateforme d'annotation automatique de catégories sémantiques. applications à la catégorisation des citations en français et en arabe. Master's thesis, Université Paris-Sorbonne, 2010. (Cité en page 41.)

[Amardeilh 2005] Florence Amardeilh, Philippe Laublet et Jean-Luc Minel. *Document annotation and ontology population from linguistic extractions.* In Proceedings of the 3rd international conference on Knowledge capture, K-CAP '05, pages 161–168, New York, NY, USA, 2005. ACM. (Cité en page 81.)

[Amblard 2007] Maxime Amblard. *Calculs de représentations sémantiques et syntaxe générative : les grammaires minimalistes catégorielles.* These, Université Sciences et Technologies - Bordeaux I, Septembre 2007. (Cité en page 35.)

[Amblard 2008] Maxime Amblard, Johannes Heinecke et Estelle Maillebuau. *Discourse Representation Theory et graphes sémantiques : formalisation sémantique en contexte industriel.* In Traitement Automatique des Langues Naturelles, Avignon, France, 2008. (Cité en page 35.)

[Amini 2001] Massih-Reza Amini. *Apprentissage Automatique et Recherche d'Information : application à l'Extraction d'Information de surface et au Résumé de Texte.* PhD thesis, Université de Paris 6, 2001. (Cité en pages xi, 22, 27 et 28.)

[Amini 2003] Massih Amini et Patrick Gallinari. *Apprentissage Numérique pour le Résumé de Texte.* In Workshop of ATALA on Automatic Summarization : solutions and perspectives, 2003. (Cité en pages v, 1, 23, 25, 27 et 28.)

[Amini 2010] Massih-Reza Amini et Cyril Goutte. *A co-classification approach to learning from multilingual corpora.* In Springer, editeur, Machine Learning Journal, pages 105–121, 2010. (Cité en page 29.)

[Amsili 2002] Pascal Amsili, Céline Raynal et Laurent Roussarie. *Stop Presupposing the Computation of Presuppositions : The Case of the French Adjective seul.* In Workshop on Information Structure in Context, pages 86–97, Stuttgart, 2002. (Cité en page 30.)

[Amsili 2004] Pascal Amsili et Laurent Roussarie. *Vers une lambda-DRT étendue.* In Actes de l'atelier sur la SDRT à TALN 2004 (11ème Conférence sur le Traitement Automatique des Langues Naturelles), Avril 2004. (Cité en page 35.)

[Asher 1993] Nicholas Asher. Reference to abstract objects in discourse. SLAP 50, Dordrecht, Kluwer, 1993. (Cité en pages 30, 35 et 36.)

[Asher 2000] Nicholas Asher. *Computation and Storage in Discourse Interpretation*, 2000. (Cité en page 30.)

[Asher 2001] Nicholas Asher, Daniel Hardt et Joan Busquets. *Discourse Parallelism, Ellipsis, and Ambiguity.* Journal of Semantics, vol. 18, no. 1, pages 1–25, 2001. (Cité en pages 35, 37 et 38.)

[Asher 2003] Nicholas Asher et Alex Lascarides. Logics of conversation. Cambridge :Cambridge University Press., 2003. (Cité en pages xiii et 67.)

[Aït El Mekki 2004] T. Aït El Mekki et A. Nazarenko. *L'index de fin de livre, une forme de résumé indicatif ?* Revue Traitement Automatique des Langues, vol. 1, no. 45, pages 121–150, 2004. (Cité en page 12.)

[Azmi 2012] Aqil M. Azmi et Suha Al-Thanyyan. *A text summarizer for Arabic.* Computer Speech &amp ; Language, vol. 26, no. 4, pages 260 – 273, 2012. (Cité en page 1.)

[Baccour 2004] Leila Baccour. *Conception et réalisation d'un système de segmentation de textes arabes non voyellés.* PhD thesis, Faculté des Sciences Économiques et de Gestion de Sfax, 2004. (Cité en page 54.)

[Ballard 1971] D. Lee Ballard, Robert J. Conrad et Robert E. Longacre. More on the deep and surface grammar of interclausal relations. Summer Institute of Linguistics, Santa Ana, CA, 1971. (Cité en page 64.)

[Belguith 1999] Lamia Hadrich Belguith. Traitement des erreurs d'accord de l'arabe basé sur une analyse syntagmatique étendue pour la vérification et une analyse multicritère pour la correction. Thèse, Faculté des Sciences de Tunis, 1999. (Cité en pages 49, 56, 58 et 82.)

[Belguith 2005] Lamia Hadrich Belguith, Leila Baccour et Mourad Ghassan. *Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules.* Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles TALN'2005 - Dourdan France, vol. Vol. 1, pages 451–456, 2005. (Cité en pages xiii, 50, 51, 54, 55, 56, 82, 96, 99 et 122.)

[Belguith 2006] Lamia Hadrich Belguith et Nouha Chaaben. *Analyse et désambiguïsation morphologiques de textes arabes non voyellés.* In 13ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2006), pages 493–501, Leuven- Belgique, 10-13 avril 2006 2006. (Cité en pages 50, 52 et 100.)

[Belguith 2008] Lamia Hadrich Belguith, Chafik Aloulou et Abdelmajid Ben Hamadou. *MASPAR : De la segmentation à l'analyse syntaxique de textes arabes.* In CÉPADUÈS-Editions, editeur, Revue Information Interaction Intelligence I3, volume 7, pages 9 – 36, http ://www.revue-i3.org/, mai 2008. 2008. ISSN : 1630-649x. (Cité en pages 51, 59 et 82.)

[Belguith 2009] Lamia Hadrich Belguith. *Analyse et résumé automatiques de documents : Problèmes, conception et réalisation.* Habilitation universitaire en informatique, Faculté des Sciences Économiques et de Gestion de Sfax (FSEG-SFAX), 2009. (Cité en page 53.)

[Belguith 2013] Lamia Hadrich Belguith, Mariem Ellouze, Mohamed Hedi Maaloul, Maher Jaoua, Fatma Kallel Jaoua et Philippe Blache. *Automatic summarization of Semitic languages.* Book Natural Language Processing for Semitic Languages - Springer, 2013. (Cité en page 125.)

[Ben Hamadou 1995] Ambedlmajid Ben Hamadou, Emna Ben Mefteh et Maher Jaoua. *Une méthode d'extraction des idées clés d'un document en vue de le résumer.* In IA'95-Quinzième journées internationales de génie linguistique, 1995. (Cité en page 20.)

[Benajiba 2009] Yassine Benajiba, Mona T. Diab et Paolo Rosso. *Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition.* Int. Arab J. Inf. Technol., vol. 6, no. 5, pages 463–471, 2009. (Cité en page 60.)

[Berri 1996] Jawad Berri. *Mise en oeuvre de la méthode d'exploration contextuelle pour le résumé automatique de textes. Implémentation du système SERAPHIN.* In Actes du colloque CLIM'96, Montréal, Canada, JUN 1996. (Cité en pages 40 et 41.)

[Bikel 2004a] Daniel M. Bikel. *Intricacies of Collins' Parsing Model.* Computational Linguistics, vol. 30, no. 4, pages 479–511, 2004. (Cité en page 56.)

[Bikel 2004b] Daniel M. Bikel. On the parameter space of generative lexicalized statistical parsing models. Master's thesis, University of Pennsylvania, 2004. (Cité en page 57.)

[Blais 2008] Antoine Blais. Résumé automatique de textes scientifiques et construction de fiches de synthèse catégorisées : Approche linguistique par annotations sémantiques et réalisation informatique. Master's thesis, UNIVERSITE PARIS IV-SORBONNE, 2008. (Cité en pages 7, 9 et 11.)

[Bossard 2010] Aurélien Bossard, Michel Généreux et Thierry Poibeau. *Résumé automatique de textes d'opinion.* Traitement Automatique des Langues, vol. 51, no. 3, pages 47–73, 2010. (Cité en page 12.)

[Bourigault 2000] D Bourigault et C Fabre. *Approche linguistique pour l'analyse syntaxique de corpus.* Cahiers de Grammaire - Université Toulouse le Mirail, no. 25, pages 131–151, 2000. (Cité en page 56.)

[Buckwalter 2004] Tim Buckwalter. *Issues in Arabic orthography and morphology analysis.* In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Semitic '04, pages 31–34, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. (Cité en page 59.)

[Busquets 2001] Joan Busquets, Laure Vieu et Nicholas Asher. *La SDRT : une approche de la cohérence du discours dans la tradition de la sémantique dynamique.* Verbum, vol. XXIII, pages 73–101, 2001. (Cité en pages xi, 37, 38 et 39.)

[Carletta 1996] Jean Carletta. *Assessing Agreement on Classification Tasks : The Kappa Statistic.* Computational Linguistics, vol. 22, no. 2, pages 249–254, 1996. (Cité en page 115.)

[Chaaben 2004] Nouha Chaaben et Lamia Belguith. *Implémentation du système MORPH2 d'analyse morphologique pour l'arabe non voyellé.* Quatrièmes journées scientifiques des jeunes chercheurs en Génie Electrique et Informatique (GEI'2004), 2004. (Cité en page 59.)

[Chaaben 2010] Nouha Chaaben, Lamia Hadrich Belguith et Abdelmajid Ben Hamadou. *The MORPH2 new version : A robust morphological analyzer for Arabic texts.* In Actes des 10èmes journées internationales d'analyse statistique des données JADT'2010, Rome, Italy, 2010. (Cité en pages 52, 96 et 122.)

[Chali 2009] Yllias Chali, Sadid Hasan et Shafiq Joty. *A SVM-Based Ensemble Approach to Multi-Document Summarization.* In Yong Gao et Nathalie Japkowicz, editeurs, Advances

in Artificial Intelligence, volume 5549 of *Lecture Notes in Computer Science*, pages 199–202. Springer Berlin / Heidelberg, 2009. (Cité en page 80.)

[Charolles 1989] Michel Charolles et André Petitjean. Le résumé de texte : aspects linguistiques, sémiotiques, psycholinguistiques et automatiques. Collogue International de Linguistique, 1989. (Cité en page 12.)

[Chiang 2000] David Chiang. *Statistical parsing with an automatically-extracted tree adjoining grammar*. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, ACL '00, pages 456–463, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. (Cité en page 57.)

[Chiang 2006] David Chiang, Mona T. Diab, Nizar Habash, Owen Rambow et Safiullah Shareef. *Parsing Arabic Dialects*. In EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy. The Association for Computer Linguistics, 2006. (Cité en page 57.)

[Chiraz 2003] zribi Chiraz Ben Othmane et Mohamed Ben Ahmed. *Efficient Automatic Correction of Misspelled Arabic Words Based on Contextual Information*. In KES, pages 770–777, 2003. (Cité en page 56.)

[Chomsky 1959] Noam Chomsky. *On Certain Formal Properties of Grammars*. Information and Control, vol. 2, no. 2, pages 137–167, 1959. (Cité en page 31.)

[Chomsky 2007] Noam Chomsky. *SymposiumonMargaret Boden, Mind as Machine : A History of Cognitive Science , Oxford (2006) two volumes*. Artif. Intell., vol. 171, no. 18, pages 1094–1103, 2007. (Cité en page 31.)

[Coates-Stephens 1992] Sam Coates-Stephens. *The Analysis and Acquisition of Proper Names for the Understanding of Free Text*. Computers and the Humanities, vol. 26, pages 441–456, 1992. 10.1007/BF00136985. (Cité en page 60.)

[Cohen 1960] J. Cohen. *A Coefficient of Agreement for Nominal Scales*. Educational and Psychological Measurement, vol. 20, no. 1, page 37, 1960. (Cité en page 115.)

[Cole 1996] Ronald A Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen et Victor Zue. Survey of the state of the art in human language technology. CSLU, Oregon Graduate Institute, 1996. (Cité en page 110.)

[Collectif 2009] Collectif. Le petit larousse illustré 2010. Le Petit Larousse Illustre. Laurier Books Ltd, 2009. (Cité en page 8.)

[Collins 2005] Michael Collins et Terry Koo. *Discriminative Reranking for Natural Language Parsing*. Comput. Linguist., vol. 31, no. 1, pages 25–70, Mars 2005. (Cité en page 56.)

[Corblin 2001] Francis Corblin et Marie-Christine Laborde. *Anaphore nominale et référence mentionnelle :* le premier, le second, l'un *et* l'autre. In W. de Mulder, C. Vet et C. Vet-

ters, editeurs, Anaphores pronominales et nominales. Etudes pragma-sémantiques, pages 99–121. Rodopi, Amsterdam, 2001. (Cité en pages 35 et 36.)

[Crabbé 2007] Benoît Crabbé. *Problématique de conception d'un langage de haut niveau.* In Atelier Formalismes Syntaxiques de Haut Niveau - TALN 2007, Toulouse, France, 2007. (Cité en page 34.)

[Crispino 2003] Gustavo Crispino. *Une plate-forme informatique de l'Exploration Contextuelle : modélisation, architecture et réalisation (ContextO) - Application au filtrage sémantique de textes.* PhD thesis, UNIVERSITE DE PARIS IV – SORBONNE, 2003. (Cité en page 40.)

[Daelemans 2003] Walter Daelemans, Véronique Hoste, Fien De Meulder et Bart Naudts. *Combined Optimization of Feature Selection and Algorithm Parameters in Machine Learning of Language.* In Nada Lavrac, Dragan Gamberger, Ljupco Todorovski et Hendrik Blockeel, editeurs, ECML, volume 2837 of *Lecture Notes in Computer Science*, pages 84–95. Springer, 2003. (Cité en page 60.)

[Debili 1998] Fathi Debili et Emna Souissi. *Etiquetage grammatical de l'arabe voyell ou non.* In Proceedings of the Workshop on Computational Approaches to Semitic Languages, Semitic '98, pages 16–25, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics. (Cité en pages 52 et 58.)

[Debili 2002] Fathi Debili, Hadhémi Achour et Emna Souissi. *De l'étiquetage grammatical à la voyellation automatique de l'arabe.* In in Correspondances, volume vol 71, pages 10–26, IRMC : Institut de Recherche sur le Maghreb Contemporain, Tunis, 2002. (Cité en pages xiii, 49, 52, 53 et 58.)

[Dempster 1977] A. P. Dempster, N. M. Laird et D. B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, vol. 39, no. 1, pages 1–38, 1977. (Cité en page 27.)

[Desclés 2006] Jean-Pierre Desclés. *Contextual Exploration Processing for Discourse and Automatic Annotations of Texts.* In FLAIRS Conference, pages 281–284, 2006. (Cité en page 40.)

[Desclés 1997] Jean-Pierre Desclés. *Systèmes d'exploration contextuelle.* In Co-texte et calcul du sens - (Claude Guimier), pages 215–232, Presses de l'universitaires de Caen, 1997. (Cité en pages 40 et 54.)

[Djioua 2007] Brahim Djioua et Jean-Pierre Desclés. *Indexing Documents by Discourse and Semantic Contents from Automatic Annotations of Texts.* In FLAIRS Conference, pages 356–361, 2007. (Cité en page 40.)

[Douzidia 2004a] Fouad Douzidia. Résumé automatique de texte arabe. Master's thesis, Université de Montréal, nov 2004. (Cité en pages 1, 20 et 21.)

[Douzidia 2004b] Fouad Douzidia et Guy Lapalme. *Lakhas, an Arabic summarization system.* In Proceedings of DUC'04, pages 128–135, Boston, may 2004. NIST, NIST. (Cité en page 2.)

[Edmundson 1969] H. P. Edmundson. *New methods in Automatic Extracting.* Journal of the Association for Computing Machinery, vol. 16(2), no. 2, pages 264 – 285, April 1969. (Cité en pages 2, 20 et 21.)

[El-Haj 2011] Mahmoud El-Haj, Udo Kruschwitzc et Chris Fox. *Experimenting with automatic text summarisation for arabic.* In Proceedings of the 4th conference on Human language technology : challenges for computer science and linguistics, LTC'09, pages 490–499, Berlin, Heidelberg, 2011. Springer-Verlag. (Cité en page 1.)

[Ellouze 2004] Mariem Ellouze. *Des schémas rhétoriques pour le contrôle de la cohérence et génération de résumés automatiques d'articles scientifiques.* PhD thesis, Université de Manouba, Ecole Nationale des sciences de l'Informatique, 2004. (Cité en pages 19, 20 et 21.)

[Estratat 2004] Mathieu Estratat et Laurent Henocque. *Reconnaître des langages avec un configurateur.* JNPC'04, 2004. (Cité en page 35.)

[Gabbard 2008] Ryan Gabbard et Seth Kulick. *Construct State Modification in the Arabic Treebank.* In Proceedings of ACL-08 : HLT, Short Papers, pages 209–212, Columbus, Ohio, June 2008. Association for Computational Linguistics. (Cité en page 57.)

[Gamut 1991] L. T. F. Gamut. Logic, language, and meaning. The University of Chicago Press, 1991. (Cité en pages 31 et 34.)

[Ghassan 2001] Mourad Ghassan. *Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des Applications informatiques : SegATex et CitaRE.* PhD thesis, Paris-Sorbonne, 2001. (Cité en page 54.)

[Giquel 1990] Françoise Giquel. Réussir le résumé de texte. Méthod'sup, 1990. (Cité en pages xi, 9 et 10.)

[Grize 1990] J.-B. Grize. *Résumer, mais pour qui ?* "Le resumé de texte : Aspects linguistiques, sémiotiques, psycholinguistiques, didactiques et automatiques", Abbaye des Prémontres, Pont-à-Mousson, 1990. (Cité en pages 12 et 14.)

[Grosz 1986] Barbara J. Grosz et Candace L. Sidner. *Attention, intentions, and the structure of discourse.* Comput. Linguist., vol. 12, no. 3, pages 175–204, Juillet 1986. (Cité en page 30.)

[Hahn 1998a] Udo Hahn. *Automatic Extracting - A Poor Man's Approach to Automatic Abstracting.* International Workshop on Extraction, Filtering and Automatic Summarization (RIFRA' 98), 1998. (Cité en pages 8 et 21.)

[Hahn 1998b] Udo Hahn et Inderjeet Mani. *Tutorial T6 : Automatic Text Summarisation.* ECAI'98, 1998. (Cité en page 8.)

[Halliday 1976] M. A. K. Halliday et Ruqaiya Hasan. Cohesion in English (English Language). Longman Pub Group, Mai 1976. (Cité en page 64.)

[Hasler 2007] Laura Hasler. *From extracts to abstracts : Human summary production operations for computer-aided summarisation.* PhD thesis, School of Humanities, Languages and Social Sciences, University of Wolverhampton, Wolverhampton, UK, June 2007. (Cité en page 12.)

[Hobbs 1978] Jerry R. Hobbs. *Coherence and Coreference.* Rapport technique 168, AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, Aug 1978. (Cité en page 30.)

[Hovy 1998] Eduard Hovy et Chin-Yew Lin. *Automated text summarization and the SUMMARIST system.* In Proceedings of a workshop on held at Baltimore, Maryland : October 13-15, 1998, TIPSTER '98, pages 197–214, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics. (Cité en page 64.)

[Hovy 2005] Eduard Hovy, Chin yew Lin et Liang Zhou. *Evaluating DUC 2005 using Basic Elements.* In Proceedings of DUC-2005, 2005. (Cité en page 112.)

[Hovy 2006] Eduard Hovy, Chin yew Lin, Liang Zhou et Junichi Fukumoto. *Automated Summarization Evaluation with Basic Elements.* In Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC'06), Genoa, Italie, 24–26 mai 2006 2006. (Cité en page 112.)

[Ide 1996] Nancy Ide et Jean Veronis. *Une application de la TEI aux industries de la langue : le Corpus Encoding Standard.* In Cahiers GUTenberg n 24 (spécial TEI), pages 166–169. 1996. (Cité en page 65.)

[Ishikawa 2001] Kai Ishikawa, Shinichi Ando et Akitoshi Okumura. *Hybrid Text Summarization Method based on the TF Method and the LEAD Method.* In In Proceedings of the 2nd National Institute of Informatics Test Collection Information Retrieval (NTCIR) Workshop, pages 5–219, 2001. (Cité en page 20.)

[Jaoua 2011] Fatma Kallel Jaoua. Une méthode flexible de résumé automatique de documents multiples. Thèse en informatique, Faculté des Sciences Economiques et de Gestion de Sfax., 2011. (Cité en page 84.)

[Jing 1998] H. Jing, R. Barzilay, K. McKeown et M. Elhadad. *Summarization evaluation methods : experiments and analysis.* In Proceedings of the Spring Symposium on Intelligent Text Summarization (AAAI 98), pages 60–68, Stanford, CA, March 1998. (Cité en page 110.)

[Jones 1993] Karen Sparck Jones. *What Might be in a Summary ?* In Information Retrieval, pages 9–26, 1993. (Cité en page 11.)

[Jones 1996] Karen Sparck Jones et Julia R. Galliers. Evaluating natural language processing systems : An analysis and review. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. (Cité en page 109.)

[Kamp 1981] Hans Kamp. *Référence Temporelle et Représentation Du Discours.* Language, vol. 64, pages 39–64, 1981. (Cité en page 34.)

[Kamp 1988] Hans Kamp. *Discourse Representation Theory : What it is and Where it Ought to Go.* In Natural Language at the Computer, pages 84–111, 1988. (Cité en pages 30, 31 et 34.)

[Kamp 1993] Hans Kamp. *Disambiguation in discourse.* In M. Aurnague et M. Bras, editeurs, Proceedings of the Fourth Toulouse Workshop on the Verbalization of Space, Time and Motion, 1993. (Cité en page 34.)

[Keerthi 2001] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya et K.R.K. Murthy. *Improvements to Platt's SMO Algorithm for SVM Classifier Design.* Neural Computation, vol. 13, no. 3, pages 637–649, 2001. (Cité en page 89.)

[Kianmehr 2009] Keivan Kianmehr, Shang Gao, Jawad Attari, M. Mushfiqur Rahman, Kofi Akomeah, Reda Alhajj, Jon Rokne et Ken Barker. *Text summarization techniques : SVM versus neural networks.* In Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services, iiWAS '09, pages 487–491, New York, NY, USA, 2009. ACM. (Cité en page 80.)

[Klein 2003] Dan Klein et Christopher D. Manning. *Accurate Unlexicalized Parsing.* In 41st Annual Meeting of the Association for Computational Linguistics (ACL'2003), pages 423–430, 2003. (Cité en page 57.)

[Kulick 2009] Seth Kulick et Ann Bies. *Treebank Analysis and Search Using an Extracted Tree Grammar*, 2009. (Cité en page 57.)

[Kupiec 1995] Julian Kupiec, Jan Pedersen et Francine Chen. *A trainable document summarizer.* In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '95, pages 68–73, New York, NY, USA, 1995. ACM. (Cité en page 25.)

[Landis 1977] J R Landis et G G Koch. *The measurement of observer agreement for categorical data.* Biometrics, vol. 33, no. 1, pages 159–174, 1977. (Cité en pages xiv et 116.)

[Lascarides 1993] Alex Lascarides et Nicholas Asher. *Temporal interpretation, discourse relations and commonsense entailment.* Linguistics and Philosophy, vol. 16, pages 437–493, 1993. 10.1007/BF00986208. (Cité en page 36.)

[Lascarides 2007] Alex Lascarides et Nicholas Asher. *Segmented Discourse Representation Theory : Dynamic Semantics With Discourse Structure*. In Harry Bunt et Reinhard Muskens, editeurs, Computing Meaning, volume 83, chapitre 5, pages 87–124. Springer Netherlands, Dordrecht, 2007. (Cité en page 30.)

[Lascarides 2009] Alex Lascarides et Nicholas Asher. *Agreement, Disputes and Commitments in Dialogue*. J Semantics, page ffn013, Fier 2009. (Cité en page 36.)

[Le Priol 2009] F. Le Priol, M. Bertin et A. Blais. Annotations automatiques et recherche d'informations. Hermes, traite ic2 – serie cognition et traitement de línformation édition, 2009. (Cité en page 40.)

[Lehmam 1995] Abderrafih Lehmam. Le résumé automatique des textes scientifiques et techniques ; aspects linguistiques et computationnels : réalisation du système rafi- (résumé automatique par fragments indicateurs). Master's thesis, Université de Nancy 2, 1995. (Cité en pages vi et 2.)

[Li 2007] Sujian Li, You Ouyang, Wei Wang et Bin Sun. *Multi-document Summarization Using Support Vector Regression*, 2007. (Cité en page 88.)

[Lin 1998] Chin Y. Lin. *Assembly of Topic Extraction Modules in SUMMARIST*. In AAAI Spring Symposium on Intelligent Text Summarisation, Stanford, California, USA, Mars 1998. (Cité en page 25.)

[Lin 2004] Chin-Yew Lin. *ROUGE : A Package for Automatic Evaluation of Summaries*. In Stan Szpakowicz Marie-Francine Moens, editeur, Text Summarization Branches Out : Proceedings of the ACL-04 Workshop, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. (Cité en page 112.)

[Longacre 1979] Robert Longacre. *The Paragraph as a Grammatical Unit*. In T. Givon, editeur, Syntax and Semantics 12. Academic Press, 1979. (Cité en page 64.)

[Luc 2001] Christophe Luc. *Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte*. TALN'2001 - Université de Tours, pages 263–272, 2001. (Cité en pages 31, 32 et 66.)

[Luhn 1958] H.P. Luhn. *The Automatic Creation of Literature Abstracts*. IBM Journal, vol. 2, pages 159–165, 1958. (Cité en pages 7, 18 et 19.)

[Maaloul 2010a] Mohamed Hédi Maaloul et Iskandar keskes. *Résumé automatique de documents arabes basé sur la technique RST*. In 12èmes Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (JEP-TALN-RECITAL'2010), Québec, Canada, 2010. (Cité en pages vi, 64, 66, 76, 78, 114, 117 et 125.)

[Maaloul 2010b] Mohamed Hédi Maaloul, Iskander keskes et Lamia Hadrich Belguith. التلخيص الآلي لِلنصوص العربية اعتمَادًا علَى نظرية البنية البلَاغية. In 6th International Computing Conference in Arabic 2010 (ICCA'2010), Hammamet, Tunisia, 2010. Cet article

a obtenu le prix du meilleur article et de la meilleure présentation. (Cité en pages 71, 84 et 126.)

[Maaloul 2010c] Mohamed Hédi Maaloul, Iskander Keskes, Lamia Hadrich Belguith et Philippe Blache. *Automatic summarization of arabic texts based on RST technique*. In 12th International Conference on Enterprise Information Systems (ICEIS'2010), Funchal, Madeira - Portugal, 2010. (Cité en page 126.)

[Maaloul 2010d] Mohamed Hédi Maaloul, Mohamed Mahdi Boudabous et Lamia Hadrich Belguith. *Digital learning for summarizing Arabic documents*. In Proceedings of the 7th international conference on Advances in natural language processing, IceTAL'10, pages 79–84, Berlin, Heidelberg, 2010. Springer-Verlag. (Cité en pages 65, 69 et 126.)

[Maaloul 2011] Mohamed Hédi Maaloul, Keskes Iskander, Boudabous Mohamed Mahdi et Hadrich Belguith Lamia. التلخيص الآلي لِلنصوص العربية : بين نظرية التلقين الرقمي و نظرية البنية البلَاغية. In 7th International Computing Conference in Arabic 2011 (ICCA'2011), Riyadh, Saudi Arabia, 2011. (Cité en pages xi, 74, 75, 81 et 125.)

[Maaloul 2012a] Mohamed Hedi Maaloul, Wajdi Ajjel et Lamia Hadrich Belguith. دور التحليل الـلُغوي في رصد العلَاقَات البلَاغية. In International Conference on Arabic Language Processing., Rabat, Morocco, 2012. CITALA'2012. (Cité en pages 114 et 125.)

[Maaloul 2012b] Mohamed Hédi Maaloul, Iskandar Keskes, Mohamed Mahdi Boudabous et Lamia Hadrich Belguith. *Étude comparative entre trois approches de résumé automatique de documents arabes (Comparative Study of Three Approaches to Automatic Summarization of Arabic Documents) [in French]*. In Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2 : TALN, pages 225–238, Grenoble, France, June 2012. ATALA/AFCP. (Cité en page 125.)

[Maamouri 2008] Mohamed Maamouri, Ann Bies et Seth Kulick. *Enhancing the Arabic treebank : a collaborative effort toward new annotation guidelines*. In In LREC'08, 2008. (Cité en page 57.)

[Makram Boujelben 2008] Chafik Aloulou et Lamia Hadrich Belguith Makram Boujelben. *Toward a detection/correction system for the agreement errors in non-voweled Arabic texts*. In The 9th International Arab Conference on Information Technology ACIT'2008, 2008. (Cité en page 58.)

[Maâloul 2008] Mohamed Hédi Maâloul, Mariem Ellouze Khemakhem et Lamia Hadrich Belguith. *Al Lakas El'eli* - الـلَخَاص الآلي *: Un système de résumé automatique de documents arabes*. In International Business Information Management Association (IBIMA'2008), Maroc, 04 au 06 janvier 2008 2008. (Cité en pages 80, 85 et 90.)

[Manguin 2003] Jean-Luc Manguin. *Utilisation d'un corpus catégorisé pour l'étude et la représentation de la synonymie en contexte.* In Actes des 3èmes journées de Linguistique de corpus, Lorient, France, 2003. (Cité en page 65.)

[Mani 1998] Inderjeet Mani et Eric Bloedorn. *Machine Learning of Generic and User-Focused Summarization.* In Proceedings of the Fifthteen National Conference on Artificial Intelligence, pages 821 – 826, Madison, Wisconsin, 1998. MIT Press. (Cité en pages xiii, 24 et 25.)

[Mani 1999] Inderjeet Mani. Advances in automatic text summarization. MIT Press, Cambridge, MA, USA, 1999. (Cité en page 8.)

[Mani 2001a] Inderjeet Mani. Automatic summarization. Natural Language Processing. John Benjamins Publishing Company, 2001. (Cité en page 8.)

[Mani 2001b] Inderjeet Mani et Mark T. Maybury. *Automatic Summarization.* In ACL (Companion Volume), page 5, 2001. (Cité en page 8.)

[Mann 1988] William C. Mann et Sandra A. Thompson. *Rhetorical structure theory : Toward a functional theory of text organization.* Text, vol. 8, no. 3, pages 243–281, 1988. (Cité en pages xi, 30, 31, 32, 37, 64, 73, 78, 83 et 103.)

[Marcu 1997] Daniel Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.* PhD thesis, University of Toronto, 1997. (Cité en page 31.)

[Marcu 2000a] Daniel Marcu. The theory and practice of discourse parsing and summarization. MIT Press, Cambridge, MA, USA, 2000. (Cité en page 43.)

[Marcu 2000b] Daniel Marcu, Lynn Carlson et Maki Watanabe. *The Automatic Translation of Discourse Structures.* In ANLP, pages 9–17, 2000. (Cité en pages 34 et 76.)

[Marcus 1980] M. Marcus. A theory of syntactic recognition for natural language. MIT Press, Cambridge, MA, 1980. (Cité en page 30.)

[Masson 1998] Nicolas Masson. *Méthodes pour une génération variable de résumé automatique : vers un système de réduction de texte.* PhD thesis, Université Paris 11-Orsay, 1998. (Cité en pages 8, 11 et 13.)

[Mathkour 2008] Hassan I. Mathkour, Ameur A. Touir et Waleed A. Al-sanea. *Parsing Arabic Texts Using Rhetorical Structure Theory.* Science Publications, 2008. (Cité en pages 1, 30, 64 et 67.)

[Mesfar 2008] Slim Mesfar. *Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard.* PhD thesis, Université de Franche-Comte, École doctorale «Langages, Espaces, Temps, Sociétés», 2008. (Cité en page 60.)

[Mezghani 2008] Souha Hammami Mezghani, Lamia Hadrich Belguith et Abdelmajid Ben Hamadou. *Anaphora in Arabic Language : developing a corpora annotating tool for ana-*

phoric links. In The 9th International Arab Conference on Information Technology ACIT'2008, 2008. (Cité en page 58.)

[Minel 2002a] Jean-Luc Minel. Filtrage sémantique : du résumé automatique à la fouille de textes. 2002. (Cité en pages xi, 12, 18, 25, 41, 66, 85, 113, 114 et 118.)

[Minel 2002b] Jean-Luc Minel. *Filtrage sémantique de textes. Problèmes, conception et réalisation d'une plate-forme informatique*. PhD thesis, Université Paris-Sorbonne, 2002. (Cité en pages xi, 23, 24 et 40.)

[Minel 2009] J-L. Minel, J-P. Desclés, E. Cartier, G. Crispino, S. Ben Hazez et A. Jackiewicz. *Résumé automatique par filtrage sémantique d'informations dans des textes*. Revue Techniques et Sciences Informatiques, 2009. (Cité en page 30.)

[Mitkov 1998] Ruslan Mitkov, Lamia Belguith et Malgorzata Stys. *Multilingual Robust Anaphora Resolution*. In In Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing, pages 7–16, 1998. (Cité en page 58.)

[Moeschler 2011] Jacques Moeschler. *Causal, Inferential and Temporal Connectives : Why parce que Is The Only Causal Connective in French*. Presses Universitaires de Rouen et du Havre, pages 97–114, 2011. (Cité en page 35.)

[Montague 1973] Richard Montague. *The Proper Treatment of Quantification in Ordinary English*. In Approaches to Natural Language, volume 49, pages 221–242. Dordrecht, 1973. (Cité en page 30.)

[Moore 1992] Johanna D. Moore et Martha E. Pollack. *A Problem for RST : The Need for Multi-Level Discourse Analysis*. Computational Linguistics, vol. 18, pages 537–544, 1992. (Cité en page 37.)

[Mouelhi 2008] Zoubeir Mouelhi. *AraSeg. *. : un segmenteur semi-automatique des textes arabes*. JADT'2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles., 2008. ICAR - Université Lumière-Lyon2. (Cité en page 51.)

[Nenkova 2004] Ani Nenkova et Rebecca Passonneau. *Evaluating content selection in summarization : The Pyramid method*. In Proceedings of HLT/NAACL2004, 2004. (Cité en pages 112 et 113.)

[Nenkova 2007] Ani Nenkova, Rebecca Passonneau et Kathleen McKeown. *The Pyramid Method : Incorporating human content selection variation in summarization evaluation*. ACM Trans. Speech Lang. Process., vol. 4, no. 2, Mai 2007. (Cité en page 112.)

[Nikolopoulos 1997] Chris Nikolopoulos. Expert systems : Introduction to first and second generation and hybrid knowledge based systems. Marcel Dekker, 1997. (Cité en page 22.)

[Ono 1994] Kenjl Ono, Kazuo Sumlta et Seijl Miike. *Abstract Generation based on Rhetorical Structure Extraction*. In Proceedings of the International Conference on Computational Linguistics (COLING'94), pages 344–348, Kyoto, Japon, 1994. (Cité en page 43.)

[Ono 1996] T. Ono et S. A. Thompson. *Interaction and syntax in the structure of conversational discourse.* In SpringerVerlag, editeur, Discourse Processing : an interdisciplinary perspective, pages 67–96. In E. Hovy and D. Scott, eds., 1996. (Cité en pages xiii et 32.)

[Osório 1998] Fernando Santos Osório. Un système hybride neuro-symbolique pour l'apprentissage automatique constructif. Master's thesis, L'Institut National Polytechnique de Grenoble - I.N.P.G. Laboratoire LEIBNIZ - IMAG, 1998. (Cité en page 22.)

[Pascual 1995] Elsa Pascual et Marie-Paule Péry-Woodley. La définition dans le texte, pages 65–88. Numeéro 33. Prescot, 1995. (Cité en page 66.)

[Platt 1998] John. C. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization.* In B. Schoelkopf, C. Burges et A. Smola, editeurs, Advances in Kernel Methods - Support Vector Learning. MIT Press, 1998. (Cité en page 89.)

[Reboul 1996] Anne Reboul et Jacques Moeschler. *Faut-il continuer à faire de l'analyse de discours ?* Hermès, revue de linguistique, no. 16, pages 61–92, Mars 1996. (Cité en page 35.)

[Reboul 1998a] Anne Reboul et Jacques Moeschler. *Pertinence.* In G. Houde, D. Kayser, O. Koenig, J. Proust et F. Rastier, editeurs, Vocabulaire des sciences cognitives : neurosciences cognitives, psychologie, intelligence artificielle, linguistique et philosophie de l'esprit, page 3 p. PUF, 1998. Contribution à un ouvrage. 98-R-100 || reboul98a 98-R-100 || reboul98a. (Cité en page 35.)

[Reboul 1998b] Anne Reboul et Jacques Moeschler. Pragmatique du discours. de l'interprétation de l'énoncé à l'interprétation du discours. Armand Colin, Paris, 1998. (Cité en page 35.)

[ROUX 1993] Dominique LE ROUX et Marie-Gaëlle MONTEIL. Perspectives d'automatisation de l'activité résumante : Présentation du projet seraphinprospects of automatic abstracting :presentation of the seraphin project. 1993. (Cité en page 8.)

[Saggion 2000] Horacio Saggion. *Génération automatique de résumés par analyse sélective.* PhD thesis, Département d'informatique et de recherche opérationnelle. Faculté des arts et des sciences. Université de Montréal, August 2000. (Cité en pages v, 1, 22, 32, 33 et 113.)

[Salton 1989] Gerard Salton. Automatic text processing : the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989. (Cité en page 21.)

[Salton 1997] Gerard Salton, Amit Singhal, Mandar Mitra et Chris Buckley. *Automatic text structuring and summarization.* Information Processing and Management, vol. 33, no. 3, pages 193 – 207, 1997. (Cité en page 21.)

[Simon 1983] Herbert A. Simon. *Why Should Machines Learn.* In T. Mitchell (Eds.) R. Michalski J. Carbonell, editeur, In : Machine Learning : An artificial intelligence approach,

volume Vol.1, pages 25–37. Mogan Kaufmann, San Mateo, CA - U.S.A, 1983. (Cité en page 22.)

[Sitbon 2007] Laurianne Sitbon. *Robustesse en recherche d'information : application à l'accessibilité aux personnes handicapées.* PhD thesis, Université d'Avignon, 2007. (Cité en page 30.)

[Steinwart 2008] Ingo Steinwart et Andreas Christmann. Support vector machines. Springer Publishing Company, Incorporated, 1st édition, 2008. (Cité en pages 88 et 89.)

[Stéphanie Weiser 2008] Philippe Laublet Stéphanie Weiser et Jean-Luc Minel. *Automatic Identification of Temporal Information in Tourism Web Pages.* In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair) Khalid Choukri, editeur, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http ://www.lrec-conf.org/proceedings/lrec2008/. (Cité en page 29.)

[Taboada 2006] Maite Taboada et William C. Mann. *Rhetorical Structure Theory : looking back and moving ahead.* Discourse Studies, vol. 8, pages 423–459, 2006. (Cité en page 30.)

[Teufel 1998] Simone Teufel et Marc Moens. *Sentence extraction and rhetorical classification for flexible abstracts.* In AAAI Spring Symposium on Intelligent Text Summarisation, Stanford, California, USA, Mars 1998. (Cité en pages 2, 31 et 74.)

[Torres-Moreno 2011] Juan-Manuel Torres-Moreno. Résumé automatique de documents : une approche statistique. hermes science, 2011. (Cité en pages xiv, 43, 111, 113 et 114.)

[Usunier 2006] Nicolas Usunier, Massih-Reza Amini et Patrick Gallinari. *Résumé automatique de texte avec un algorithme d'ordonnancement.* Ingénierie des Systèmes d'Information, vol. 11, no. 2, pages 71–91, 2006. (Cité en pages 25 et 26.)

[van Dijk 1983] T. A. van Dijk et W. Kintsch. Strategies of discourse comprehension. Academic Press, Inc., New York, 1983. (Cité en page 10.)

[Versteegh 2001] Kees Versteegh. The arabic language. Edinburgh University Press, 2001. (Cité en page 47.)

[Véronis 2000] Jean Véronis. *Alignement de corpus multilingues.* Paris : Éditions Hermès, pages 115–150, 2000. (Cité en page 65.)

[Wacholder 1997] Nina Wacholder, Yael Ravin et Misook Choi. *Disambiguation of proper names in text.* In Proceedings of the fifth conference on Applied natural language processing, pages 202–208, Morristown, NJ, USA, 1997. Association for Computational Linguistics. (Cité en page 60.)

[Witten 2005] Ian H. Witten et Eibe Frank. Data mining : Practical machine learning tools and techniques. Morgan Kaufmann Series in Data Management Sys. Morgan Kaufmann, second édition, June 2005. (Cité en page 89.)

[Young-Min Kim 2010] Massih-Reza Amini Young-Min Kim Jean-François Pessiot et Patrick Gallinari. *Apprentissage d'un Espace de Concepts de Mots pour une Nouvelle Représentation des Données Textuelles*. Document Numérique, vol. 13, no. 1, pages 63–82, 2010. (Cité en page 22.)

[Zribi 2010] Inès Zribi, Souha Mezghani Hammami et Lamia Hadrich Belguith. *L'apport d'une approche hybride pour la reconnaissance des entités nommées en langue arabe*. TALN'2010, 2010. (Cité en page 59.)

Annexe A

Relations rhétoriques

Dans cette annexe, nous allons présenter l'ensemble des relations rhétoriques utilisées dans l'étape d'analyse rhétorique. Ces relations rhétoriques ont pour objectif de lier deux segments de texte adjacents entre eux, dont l'un possède le statut de noyau - segment de texte primordial pour la cohérence et exprime ce qui est plus essentiel au but de l'auteur - et l'autre celui de satellite - segment optionnel.
La reconnaissance d'une relation se fait généralement par l'application d'un ensemble de contraintes sur le noyau, le satellite et sur la combinaison du noyau et du satellite.
Rappelons que ces contraintes sont regroupées dans des règles rhétoriques (frames rhétoriques) formées par des signaux linguistiques et des heuristiques observés qui sont principalement des marqueurs indépendants d'un domaine particulier mais qui ont des valeurs importantes dans un article de presse.
Les frames présentés ci-dessous sont énumérés selon leurs appartenance à une relation rhétorique.

Relation rhétorique "Condition - شرط"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "لَا"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "من دون"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ولو"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "لَا"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "دون"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Relation rhétorique "Concession - استدرَاك"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "بعض"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "لَا سيمَا"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "و إِن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "بل"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "أَمَا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "قد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنَّا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنّي"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنهم"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "أمّا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنه"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "فقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنَّا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "فقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنّي"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "فقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنهم"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "فقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "أمَا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "فقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنه"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "فقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ولكن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "لقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ولكن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "قد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "قد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنَّا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "قد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنّي"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "قد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنهم"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "قد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "أمّا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "أَمَا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "قد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنه"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "قد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ولكن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "فقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "إِلَا أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "إِلَا أَنه"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "إِلَا أَنهَا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "إِلَا أَنهم"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur " غير أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "غير أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "غير أَنه"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "غير أَنّه"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "غير أَنَّهَا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "لقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ولكن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "لقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنه"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "لقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "أَمَا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "لقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكنهم"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وقد"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ولكن"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "أَمَا"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "إِن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "أَمَا"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "لكن"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "لكنه"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "ليس لأَن"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	(A) - (أ) et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس لأَن"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	(A) - (أ) et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس لأنه"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	(A) - (أ) et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس لأنهم"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	(A) - (أ) et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس لأنهمَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	(A) - (أ) et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس لأنهَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	(A) - (أ) et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس لأنك"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	(A) - (أ) et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس لأنكم"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	(A) - (أ) et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس لأَنكمَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	(A) - (أ) et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس لأَنكن"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	(A) - (أ) et "(B) - (ب)"

Relation rhétorique "Énumération - تفصيل"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "فهذَا"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	(A) - (أ)

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "كذلك"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "و"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "وكذلك"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "و"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "وكذلك"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "هذَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "هذَا"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "كذلك"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "و"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "كمَا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "و"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "إِضَافة"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "و"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "إِضَافة إِلَى"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "و إِضَافة إِلَى"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "هذه"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "و إِضَافة إِلَى"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "هذَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Relation rhétorique "Exception - استثناء"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "هذَا"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لم"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "هذَا"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "سوَى"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Relation rhétorique "Confirmation - توكيد"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "رغم"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "و إنمَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "علَى رغم"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "ليس"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس إِلَّا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "لَا"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس إِلَّا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "لَا"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "إِلَّا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "ليس"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "إِلَّا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "بل"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لَا سيمَا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "لقد"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "إِن"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "إِنهَا"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "فَأَن"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "أَنه"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وَاذ"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "فَان"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وَاذ"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "إِن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "انهم"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "ان"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "انه"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "انهَا"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "إنمَا"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "انهم"
Contrainte sur (B) - (ب) :	-
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "أكد"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "أن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "وأكد"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "أن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Relation rhétorique "Réduction - تقليل"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ربما"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "وربما"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "فربما"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Relation rhétorique "Joint - ربط"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ثم"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "كمَا أَن"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "هذَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "كمَا أَن"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "هذه"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لكي"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "إِمَا"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "أَو"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "لَا"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "ولَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "لم يمض علَى"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "حتَى"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Relation rhétorique "Évidence - قَاعدة"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "فئِن"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "إِلَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Relation rhétorique "Négation - نفي"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "دون"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)" et "(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "ولكن"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليست"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "ولكن"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لم"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "ولكن"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ولم"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "ولكن"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "ولكن"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليس"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "ولكن"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "ليسوُا"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Relation rhétorique "Exemplification - تمثيل"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "مثل"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "و"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "مثلاً"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "و"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "فمثلاً"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "و"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Relation rhétorique "Explication - تفسير"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "لأن"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "أي"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "و"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "لذلك"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "بينما"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "لذلك"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "أما"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

Relation rhétorique "Classement - ترتيب"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وتلتهَا"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur " ثم"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "وتلتهَا"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "وتلتهَا" + contient l'indice déclencheur " ثم"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

Relation rhétorique "Conclusion - استنتَاج"

Contrainte sur (A) - (أ) :	-
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "فلذَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

Relation rhétorique "Affirmation - جزم"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "ويجزم"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "يجزم"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Relation rhétorique "Définition - تعريف"

Contrainte sur (A) - (أ) :	Contient l'indice complémentaire "لِئن"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "فهي"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(A) - (أ)"

Relation rhétorique "Pondération - ترجيح"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "لعل"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "ولعل"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "ولعله"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "ولعلهَا"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "ولعلهم"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "لعله"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "لعلهَا"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "لعلهم"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "ربمَا"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "وربمَا"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "وَالأَرجح"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "الأَرجح"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Début
Unité minimale retenue :	"(B) - (ب)"

Relation rhétorique "Possibilité - إِمكَان"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "أَو"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "كَان"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(A) - (أ)"

Relation rhétorique "Restriction - حصر"

Contrainte sur (A) - (أ) :	contient l'indice complémentaire "مَا"
Contrainte sur (B) - (ب) :	Contient l'indice déclencheur "إِلَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Relation rhétorique "Spécification - تخصيص"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "خَاصة"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire " أَن"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "خصوصًا"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "في"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "خِص يصًا"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "فيمَا"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Relation rhétorique "Justification - تعليل"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "من أَجل"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "أَن"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "كي"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "إِلَى"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Contrainte sur (A) - (أ) :	Contient l'indice déclencheur "حتَى"
Contrainte sur (B) - (ب) :	Contient l'indice complémentaire "إِن"
Position de l'indicateur déclencheur :	Milieu
Unité minimale retenue :	"(B) - (ب)"

Annexe B

Critères d'extraction

Dans cette annexe, nous présentons l'ensemble des critères d'extraction utilisés par notre système "L.A.E - الـلَخَاص الآلي" dans la phase d'apprentissage.

- La position de la phrase dans le texte : *position_ph_texte*

Type du critère : Classe {A, B, C}

Principe : Classer les phrases selon leurs positions dans le texte.

- Classe A : Si la phrase est dans le premier tiers du texte ($P < \frac{N}{3}$),
- Classe B : Si la phrase est dans le deuxième tiers du texte ($\frac{N}{3} \leqslant P < \frac{(N*2)}{3}$)
- Classe C : Sinon

avec N : Nombre de phrases dans le texte.

- La position de la phrase dans le paragraphe : *position_ph_parag*

Type du critère : Classe {A, B, C}

Principe : Classer les phrases selon leurs positions dans le paragraphe.

- Classe A : Si la phrase est dans le premier tiers du paragraphe ($P < \frac{N}{3}$),
- Classe B : Si la phrase est dans le deuxième tiers du paragraphe ($\frac{N}{3} \leqslant P < \frac{(N*2)}{3}$)
- Classe C : Sinon

avec N : Nombre de phrases dans le paragraphe.

- Le rang de la phrase : *rang_parag_ph*

Type du critère : valeur

Principe : Égale au rang du paragraphe dans lequel la phrase est placée.

- $Valeur = \frac{rang}{N}$

avec N : Nombre de paragraphes dans le texte.

- Le rang de la phrase : *position_ph_sec*

Type du critère : Classe {A, B, C}

Principe : Classer les phrases selon leurs positions dans la section.

- Classe A : Si la phrase est dans le premier tiers de la section ($P < \frac{N}{3}$),
- Classe B : Si la phrase est dans le deuxième tiers de la section ($\frac{N}{3} \leqslant P < \frac{(N*2)}{3}$)

– Classe C : Sinon

avec N : Nombre de phrases dans le la section.

- Le rang de la phrase : *rang_sec_ph*

Type du critère : valeur

Principe : Égale au rang de la section dans laquelle la phrase est placée.

– $Valeur = \frac{rang}{N}$

avec N : Nombre de sections dans le texte.

- Mots du titre : *nb_mot_tit_ph*

Type du critère : valeur

Principe : Égale à la fréquence de mots du titre dans la phrase.

– $Valeur = \frac{NB}{N}$

avec

– NB : Nombre de mots clés significatifs dans la phrase,

– N : Nombre de mots clés significatifs dans le titre.

- Mots significatifs : *nb_mot_ph*

Type du critère : valeur

Principe : Égale à la fréquence des mots significatifs (mots non vides) dans la phrase.

– $Valeur = \frac{NB}{N}$

avec

– NB : Nombre de mots significatifs dans la phrase.

– N : Nombre de mots significatifs dans la phrase la plus longue du texte.

- Position dans le texte : *pos_ph_txt*

Type du critère : valeur

Principe : Égale à la position de phrase dans le texte.

– $Valeur = \frac{R}{N}$

avec

– R : Rang de la phrase dans le texte.

– N : Nombre de phrases dans le texte.

- Position dans le paragraphe : *pos_ph_parag*

Type du critère : valeur

Principe : Égale à la position de la phrase dans le paragraphe.

- $Valeur = \frac{R}{N}$

avec

- R : Rang de la phrase dans le paragraphe.
- N : Nombre de phrases dans le paragraphe.

- Position dans la section : *pos_ph_sec*

Type du critère : valeur

Principe : Égale à la position de phrase dans la section.

- $Valeur = \frac{R}{N}$

avec

- R : Rang de la phrase dans la section.
- N : Nombre de phrases dans la section.

- Mots bonus : *nb_exp_bonus*

Type du critère : valeur

Principe : Égale à la fréquence des expressions bonus dans la phrase.

- $Valeur = \frac{NB}{N}$

avec

- NB : Nombre des expressions bonus dans la phrase.
- N : Nombre de mots significatifs (mots non vides) dans la phrase.

- Mots stigma : *nb_exp_stigma*

Type du critère : valeur

Principe : Égale à la fréquence des expressions stigma dans la phrase.

- $Valeur = \frac{NB}{N}$

avec

- NB : Nombre des expressions stigma dans la phrase.
- N : Nombre de mots significatifs (mots non vides) dans la phrase.

- Phrase anaphorique : *mot_ph_anapho*

Type du critère : Classe {A, B}

Principe : Classer les phrases selon leurs catégories (anaphorique ou non).

- Classe A : Si la phrase commence par une expression anaphorique,

– Classe B : Sinon

- Co-occurrences lexicales : $tf\_idf\_numeric$

Type du critère : valeur

Principe : Égale à la fréquence des mots clés [1] dans la phrase.

$$Ph_{fm} = \frac{1}{Nb(m)} \sum_{t \in Ph} f(m) * \frac{\log(Nb(m))}{Ph(m)} \tag{B.1}$$

avec

– $f(m)$ est la fréquence du mot m dans la phrase Ph.

– $Ph(m)$ est le nombre de phrases dans lesquelles m apparaît.

– $Nb(m)$ est le nombre de mots dans la phrase Ph.

1. Ce sont les mots qui ont une haute fréquence d'apparition dans un texte.

Résumé

Cette thèse s'intègre dans le cadre du traitement automatique du langage naturel. La problématique du résumé automatique de documents arabes qui a été abordée, dans cette thèse, s'est cristallisée autour de deux points. Le premier point concerne les critères utilisés pour décider du contenu essentiel à extraire. Le deuxième point se focalise sur les moyens qui permettent d'exprimer le contenu essentiel extrait sous la forme d'un texte ciblant les besoins potentiels d'un utilisateur.

Afin de montrer la faisabilité de notre approche, nous avons développé le système "L.A.E", basé sur une approche hybride qui combine une analyse symbolique avec un traitement numérique.

Les résultats d'évaluation de ce système sont encourageants et prouvent la performance de l'approche hybride proposée. Ces résultats, ont montré, en premier lieu, l'applicabilité de l'approche dans le contexte de documents sans restriction quant à leur thème (Éducation, Sport, Science, Politique, Reportage, etc.), leur contenu et leur volume. Ils ont aussi montré l'importance de l'apprentissage dans la phase de classement et sélection des phrases forment l'extrait final.

Mots clés :

Résumé automatique, mono-document, théorie de la structure rhétorique, arbre RST, apprentissage, algorithme SVM, approche hybride.

Abstract

This thesis falls within the framework of Natural Language Processing. The problems of automatic summarization of Arabic documents which was approached, in this thesis, are based on two points. The first point relates to the criteria used to determine the essential content to extract. The second point focuses on the means to express the essential content extracted in the form of a text targeting the user potential needs.

In order to show the feasibility of our approach, we developed the "L.A.E" system, based on a hybrid approach which combines a symbolic analysis with a numerical processing.

The evaluation results are encouraging and prove the performance of the proposed hybrid approach.These results showed, initially, the applicability of the approach in the context of mono documents without restriction as for their topics (Education, Sport, Science, Politics, Interaction, etc), their content and their volume. They also showed the importance of the machine learning in the phase of classification and selection of the sentences forming the final extract.

Key words :

Automatic summarization, mono-document, rhetorical structure theory, RST-tree, machine learning, SVM algorithm, hybrid approach.