

Thème de mémoire

**RÉALISATION D'UN SYSTÈME DE RÉSUMÉ
AUTOMATIQUE DE TEXTES ARABES**

إعداد ملخص آلي للنصوص العربية

Pour obtenir le titre
de Magister en sciences du langage et de la communication linguistique
Option : Traitement automatique de la langue

Présenté par :
BAHLOUL Belahcene

Encadré par :
Pr. ALIMAZIGHI Zaïa & Dr. ALIANE Hassina

Devant le jury :

M^r Boukhalfa Kamel, MCA - USTHB, Président

M^{elle} Mahdaoui Latifa, MCA - USTHB, Examineur

M^{me} Alimazighi Zaïa, Professeur - USTHB, Encadreur

M^{me} Aliane Hassina, Maître de recherche - CERIST, Co-encadreur

Année universitaire : 2011/2012

■ Mes plus chaleureux remerciements...

▶ à **Monsieur Boukhalfa Kamel**. MCA, USTHB, pour m'avoir fait l'honneur d'accepter de présider mon jury.

▶ à **M^{elle} Mahdaoui Latifa**. MCA, USTHB, pour m'avoir fait l'honneur d'accepter d'être membre mon jury.

▶ à **Madame Alimazighi Zaia**. Professeur, USTHB, pour m'avoir fait l'honneur d'accepter de diriger ma thèse.

▶ à **Madame Aliane Hassina**. Maître de recherche, CERIST, pour avoir été à l'origine de ce mémoire, pour son accueil, pour avoir guidé mon travail tout au long de ce dernier, et pour sa capacité à trouver des solutions simples et élégantes à des problèmes épineux.

▶ à **Monsieur Abdelmadjid Salmi**, Professeur, Chef du département de linguistique, Université d'Alger 2, pour nous avoir aidé tout au long de notre formation,

▶ à **Tout le personnel du département de linguistique**, Faculté des lettres et des langues, Université d'Alger2.

▶ à **Tout mes enseignants aux CRSTDLA**, et particulièrement le professeur **Abderrahmane Hadj salah**, pour ses efforts, et son esprit encourageux t'envers la recherche et les chercheurs,

▶ à **Tout le personnel du CRSTDLA**,

▶ enfin, à **toutes celles et tous ceux qui m'ont apporté un soutien logistique, moral ou spirituel** pendant ces dernières années.

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Résumé

La présente étude consiste en la conception et la mise en place d'un système de résumé automatique de textes arabes. Notre travail s'inscrit dans le cadre de traitement automatique de la langue arabe, étudier les différentes caractéristiques de cette langue, étudier le corpus d'articles de presse publiés sur le web et la nécessité de développer des systèmes qui produisent automatiquement des représentations abrégées de leurs contenus. Notre système de résumé intervient dans différents niveaux de traitement, filtrage, segmentation, normalisation, calcul de poids et réduction de phrases. Notre système repose sur des techniques statistiques et linguistiques d'extraction de l'information utile, ces techniques ont déjà fait leurs preuves pour les autres langues tel que l'anglais ou le français alors pourquoi pas l'arabe.

Mots clés : Traitement automatique de l'arabe, Résumé automatique, Articles de presse web, Extraction d'informations, Techniques statistiques, Techniques linguistiques.

Summary

This study involves the design and implementation of a system for summarization of Arabic texts. Our work is part of the automatic processing of Arabic language, study the different characteristics of this language, studying the corpus of newspaper articles published on the Web and the need to develop systems that automatically abbreviated representations of their contents. Our summarization system is involved in various levels of treatment, filtering, segmentation, normalization, calculation of weight and reduction of sentences. Our system is based on linguistic and statistical techniques to extract useful information, these techniques have already proven themselves in other languages such as English or French, so why not Arabic.

Keywords : Automatic processing of Arabic, Automatic summarization, Web news articles, information extraction, Statistical techniques, Linguistic techniques.

ملخص

تتضمن دراستنا هذه تصميم و إنجاز نظام آلي لتلخيص نصوص عربية. يدخل عملنا هذا في إطار المعالجة الآلية للغة العربية ، دراسة خصائص هذه اللغة، دراسة المقالات الصحفية المنشورة عبر الأنترنت و ضرورة تطوير برامج قادرة على استخلاص المعلومات الأساسية منها بصفة آلية. إن نظام التلخيص الآلي المنجز يتدخل في العديد من مستويات المعالجة بداية من التصفية، التقسيم، التسوية، حساب المقادير إلى غاية التقليل من الجمل. إن نظامنا الآلي يعتمد أساسا على مجموعة من التقنيات الإحصائية و اللغوية لاستخراج المعلومة المهمة. هذه التقنيات التي أظهرت نتائج رائعة في لغات أخرى كالانكليزية و الفرنسية فلما لا العربية.

كلمات مفتاحية : العلاج الآلي للغة العربية، الملخص الآلي، المقالات الصحفية عبر الأنترنت، استخراج المعلومات، التقنيات الإحصائية، التقنيات اللغوية.

SOMMAIRE

Introduction générale	10
Problématique	11

ETAT DE L'ART

CHAPITRE I : INTRODUCTION AU TRAITEMENT AUTOMATIQUE DES LANGUES NATURELLES [TALN]

1. Introduction	12
2. Historique du TALN	12
3. Définition du TALN	12
3.1. L'analyse	13
3.2. La génération	13
4. Outils de TALN	14
5. Domaines de recherche de TALN	14
6. Pourquoi le TALN ?	14
7. Connaissances sur la langue	16
8. Les niveaux de traitement dans un système TALN	16
6.1. Niveau phonologique (phonétique)	17
6.2. Niveau morpholexical (morphologique)	17
6.2.1. La flexion	18
8.2.2. La dérivation	18
8.3. Niveau syntaxique	18
8.4. Niveau sémantique	18
8.5. Niveau pragmatique	18
9. Architecture d'un système TALN	19
9.1. Architecture startificationnelle (Appelée aussi séquentielle ou en série)	19
9.2. Architecture moins hiérarchisée (Appelée aussi systèmes hétérarchiques)	19
10. Problème majeur de TALN : L'ambiguïté	20
11. Le Traitement Automatique de la Langue Arabe	21
12. Le Traitement Automatique de la Langue Arabe en Algérie	21
13. Conclusion	23

CHAPITRE II : LES RÉSUMÉS AUTOMATIQUES	
1. Introduction	24
2. Historique	24
3. Définitions	24
4. Extraction vs Abstraction	25
5. Objets du résumé automatique	25
6. Types de résumés	25
7. Méthodes de résumés automatiques	25
7.1. Méthodes basées sur l'analyse linguistique du document	26
7.1.1. La méthode d'exploration contextuelle	26
7.1.2. Méthode basée sur les relations (cohésion lexicale)	28
7.2. Méthodes basée sur l'analyse statistique du document	28
7.2.1. Méthodes à base de mots clés	29
7.2.1.1. Méthode à base de Mots-clé prédéfinis	29
7.2.1.2. Méthode à base de titres	29
7.2.1.3. Méthode à base de distribution des termes	30
7.2.2. Méthode à base de position	31
7.2.3. Méthode dépendant de la longueur de phrase	31
7.2.4. Méthode à base d'expressions indicatives (cue methods)	32
7.2.5. Méthode hybride	32
8. Evaluation des résumés automatiques	34
9. Conclusion	35
CHAPITRE III : L'ARABE ET LES RÉSUMÉS AUTOMATIQUES	
1. Introduction	36
2. La langue arabe	37
2.1 Particularité de la langue arabe	37
2.2. L'écriture arabe	38
2.2.1. Les lettres de l'arabe	38
2.2.2. Les diacritiques	39
2.2.3. Le tanwin	39
2.2.4. La chadda	39
2.3. Analyse morphologique de la langue arabe	40

2.3.1. Mécanisme de dérivation	40
2.3.1. Structure des mots	41
2.3.2. Catégories des mots	42
2.3.2.1. Les noms	42
2.3.2.2. Les verbes	43
2.3.2.3. Les particules	44
3. Les problèmes du résumé automatique de textes arabes	44
3.1. Agglutination des mots	44
3.2. Absence de voyelles	45
3.3. Segmentation de texte	46
3.4. Segmentation de paragraphes	46
4. Travaux sur le résumé automatique de textes arabes	46
5. Conclusion	47

CONCEPTION DU SYSTÈME DE RESUMÉ

CHAPITRE IV : NATURE ET FORMAT DU CORPUS : LES ARTICLES DE PRESSE WEB

1. Introduction	48
2. Formes d'articles de presse	48
3. Organisation des articles de presse	49
4. La presse électronique arabe	49
4.1 Diversité des articles de presse web	50
4.2 Contenu des articles de presse web	50
5. Format d'un article de presse web	51
6. Qu'est ce que HTML	51
7. Structure générale d'une page HTML	52
7.1 L'entête <HEAD></HEAD>	52
7.2 Le corps <BODY></BODY>	53
7.2.1 Les titres dans une page <H..></H..>	54
7.2.2 Les paragraphes <P></P>	54
7.2.3 Les commentaires <!--.....-->	55
7.2.4 Les sauts de ligne 	55
7.2.5 Les lignes horizontales <HR>	55

7.2.6 Les styles physiques	55
7.2.7 Les listes 	55
7.2.8 Les autres balises HTML	55
7.3 Les balises HTML en faveur à l'extraction du texte utile	56
8. L'équation du score final	56
9. Conclusion	57

CHAPITRE V : ARCHITECTURE DU SYSTÈME DE RESUMÉ

1. Introduction	58
2. Schéma général du système de résumé	58
3. Description des différents modules du système	60
3.1. Les composants du système	60
3.1.1. Le document HTML	60
3.1.2. Anti-dictionnaire HTML	62
3.1.3. Dictionnaire HTML	62
3.1.4. Dictionnaire suffix-préfix	63
3.1.5. Anti –dictionnaire linguistique	63
3.1.6. Marqueurs linguistiques d'extraction	65
3.1.7. Marqueurs HTML	66
3.1.8. Expressions indicatives	67
3.1.9. Base de données statistiques	69
3.2. Les traitements du système	69
3.2.1. Filtrage HTML	69
3.2.2. Segmentation du document en phrases	70
3.2.3. Normalisation	71
3.2.4. Lemmatisation	71
3.2.5. Suppression de mots vides	71
3.2.5. Calcul des poids de phrases	71
3.2.6. Extraction des phrases les plus pesantes	72
3.2.7. Réduction de phrases	72
3.2.7.1. Substitution de noms	72
3.2.7.2. Suppression de parties de phrases à partir de frontières	73
3.2.7.3. Suppression des constructions de discours indirect	74

3.2.8. Reformation du document	75
4. Réalisation du système de résumé	75
5. Exécution du système de résumé	76
6. Evaluation du système de résumé	83
7. Etude comparative de notre système avec d'autres systèmes	84
8. Conclusion	86
Conclusion générale	87
Références bibliographiques	88
Annexe	92

LISTE DES TABLEAUX

Tab 1. Les lettres arabes	38
Tab 2. Variation d'écriture de la lettre ع (Ayn)	39
Tab 3. Exemple d'ambiguïté causée par l'absence de voyelles	40
Tab 4. Exemple de dérivation pour les mots كتب <i>écrire</i> et حمل <i>porter</i>	41
Tab 5. Liste des préfixes et suffixes les plus fréquents	45
Tab 6. Diversités des articles de presse arabe sur le web	50
Tab 7. Contenu des articles de presse arabe sur le web	51
Tab 8. Quelques éléments de l'anti-dictionnaire HTML	62
Tab 9. Quelques éléments du dictionnaire HTML	62
Tab 10. Quelques éléments de l'anti-dictionnaire linguistique	63
Tab 11. Liste de quelques marqueurs linguistiques d'extraction	65
Tab 12. Liste de quelques marqueurs HTML	66
Tab 13. Exemples d'expressions indicatives	67
Tab 14. Exemples de classes d'expressions indicatives	68
Tab 15. Exemples de mots vides spécifiques au corpus	70
Tab 16. Exemples de substitution de noms	72
Tab 17. Exemples de mots connecteurs	73
Tab 18. Exemples d'interprétation de mots débutant par و	73
Tab 19. Quelques modèles de motifs	74
Tab 20. Exemple de suppression de construction de discours indirect	75

Tab 21. Tableau des scores finaux	82
Tab 22. Etude comparative de notre système avec les systèmes LAKHAS et Essential-summarizer	84
Tab 23. Tableau comparatif des longueurs des résumés produits par notre système avec les systèmes LAKHAS et Essential-summarizer	85

LISTE DES FIGURES

Fig 1. Le processus d'analyse	13
Fig 2. Le processus de génération	13
Fig 3 : Niveaux de traitement du langage naturel	17
Fig 4 : Architecture stratificationnelle d'un système TALN	19
Fig 5 : Architecture d'un système TALN intégré	20
Fig 6. Les MLE organisés par domaine	27
Fig 7. Structuration de la base des MLE	27
Fig 8. Mécanisme de dérivation en arabe	40
Fig 9. Structure d'un mot arabe	41
Fig 10. Structuration du lexique arabe	42
Fig 11. Structure générale d'une page HTML	52
Fig 12. Les balises <HEAD> et <META> d'un document HTML	53
Fig 13. Schéma général du système de résumé	59
Fig 14. Aperçu d'un Document html (article de presse) sur un browser	60
Fig 15. Aperçu d'une portion du squelette d'un document html	61
Fig 16. L'interface du système de résumé	77
Fig 17. Aperçu d'un fichier filtré	78
Fig 18. Aperçu d'un fichier destiné au calcul de poids	79
Fig 19. Aperçu d'un fichier ordonné	80
Fig 20. Aperçu d'un résumé final	82