MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE D'ALGER 2
FACULTE DES LETTRES ET DES LANGUES
DEPARTEMENT DE LINGUISTIQUE

Thème de mémoire

RÉALISATION D'UN SYSTÈME DE RÉSUMÉ AUTOMATIQUE DE TEXTES ARABES

إعداد ملخِّص 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, Examinateur

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

M^{me} Aliane Hassina, Maitre 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.** Maitre 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.

بسم الله الرحمن الرحيم

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 nécessite 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

CHAPITRE II : LES RÉSUMÉS AUTOMATIQUES

CHAPITRE III : L'ARABE ET LES RÉSUMÉS AUTOMATIQUES

CONCEPTION DU SYSTÈME DE RESUMÉ

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

CHAPITRE V : ARCHITECTURE DU SYSTÈME DE RESUMÉ

LISTE DES TABLEAUX

LISTE DES FIGURES