Université de Montréal

Résumé automatique de texte arabe

par

Fouad Soufiane Douzidia

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de M.Sc

en informatique

Septembre, 2004

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Résumé automatique de texte arabe

présenté par :
Fouad Soufiane Douzidia

a été évalué par un jury composé des personnes suivantes :

Philippe Langlais, président-rapporteur
Guy Lapalme, directeur de recherche
Michel Boyer, membre du jury

Mémoire accepté le 25 octobre 2004

Résumé

La forte augmentation de texte disponible en format numérique a fait ressortir la nécessité de concevoir et de développer des outils de résumé performants dans le but de repérer et extraire l'information pertinente sous une forme abrégée. Les textes arabes ne font pas exception quant à leur disponibilité mais ils manquent d'outils de traitements automatiques.

Ce mémoire propose une méthode de production de résumés pour les textes arabes. Notre démarche méthodologique consistait à étudier : les caractéristiques de la langue arabe, un corpus de texte journalistique arabe et les techniques utilisées dans le résumé automatique. L'objectif de cette étude fut de repérer les traits caractérisant le contenu essentiel d'un article, d'identifier des marqueurs linguistiques énonçant des concepts importants et d'adapter les techniques de résumé automatique aux textes arabes.

Lakhas[1], le système de résumé automatique de textes arabes que nous avons développé est basé sur des techniques d'extraction qui ont déjà fait leurs preuves pour d'autres langues comme l'anglais.

Nous avons montré la qualité de nos résultats au moyen de deux évaluations au cours desquelles nous avons comparé les résumés produits par Lakhas avec d'autres technologies de production de résumé automatique. Grâce aux techniques de compression que nous avons introduites à Lakhas, nous avons pu montrer lors d'une compétition d'évaluation de résumé automatique, que les traductions des résumés produits par Lakhas étaient meilleurs par rapport à des résumés produits à partir de textes traduits. De plus, notre étude a fait ressortir la nécessité de travailler sur les textes dans leur langue originale au lieu de textes traduits pour l'obtention de meilleurs résumés.

Mots clés : Résumé automatique, Traitement automatique de la langue arabe, Extraction de l'information, traduction arabe.

[1] Transcription stricte de *résumer* en arabe

Abstract

The increase in availability of text in digital format accentuates the need for design and development of efficient summarizer tools to track and extract relevant information in a shortened form. Arabic texts are becoming widely available but miss tools for its automatic processing.

This master's thesis proposes a method for producing summaries for Arabic texts. We present a study of the characteristics of the Arabic language, a corpus of Arabic journalistic text and the techniques used in automatic summarization. The objective of this study was to find out the features characterizing relevant content in an article, to identify linguistic markers expressing important concepts and to adapt the techniques of automatic summarization to Arabic texts.

Lakhas[1], the system of automatic summary of Arabic texts which we developed is based on techniques of extraction which turn out to be effective for other languages such as English.

We were able to show the quality of our results by two evaluations in which we compared summaries produced by Lakhas with other technologies of production of automatic summaries. Using further techniques of compression, we have shown in a competition of evaluation of automatic summaries, that the translations of summaries produced by Lakhas were the best compared to summaries of translated texts. Furthermore, our study emphasizes the advantages of working on texts in original language instead of texts translated.

Keywords: automatic summarization, Arabic language processing, information extraction, Arabic translation.

---

[1] Corresponding roughly to *summarize* in Arabic

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

Remerciements

Je tiens d'abord à remercier mon directeur de recherche, Guy Lapalme, pour sa disponibilité, sa générosité, ses conseils et orientations qui m'ont été d'une grande aide durant la réalisation de ce travail.

Merci à Paul Over du NIST pour sa collaboration et son aide pour l'obtention des correspondances pour les textes arabes.

Je remercie également Franz Och d'ISI pour nous avoir fourni les traductions en anglais de nos résumés arabes en utilisant le système de traduction automatique d'ISI.

Merci à Philippe Langlais et Michel Boyer qui ont accepté de juger ce travail et d'en être les rapporteurs.

1 Introduction

L'objectif du traitement automatique des langues est la conception de programmes capables de traiter des données exprimées dans une langue naturelle pour lesquels plusieurs phases d'analyse (morphologique, syntaxique, sémantique et pragmatique) sont nécessaires afin d'en extraire des informations.

Avec l'avènement des documents électroniques, des quantités phénoménales d'informations sont générées. Cette montée en volume de textes nécessite la production d'outils informatiques performants dont la tâche est de trouver et d'extraire l'information pertinente sous une forme condensée.

Le Résumé de Texte Automatique semble être une bonne solution qui se trouve à la croisée de deux disciplines: traitement automatique de la langue (TAL) et recherche d'information (RI). Le Résumé de Texte Automatique consiste à produire une représentation courte d'un texte tout en conservant l'information pertinente.

L'information utile est très souvent disponible mais dans des langues différentes. Ces dernières années ont donc été marquées par des recherches sur le traitement des données textuelles multilingues. La langue arabe ne fait pas exception mais elle a été beaucoup moins étudiée au point de vue informatique que l'anglais ou le français.

Pour élaborer des systèmes de résumé de texte automatique, la plupart des chercheurs se sont basés sur des systèmes à base de connaissances linguistiques. Ces systèmes utilisent essentiellement des techniques d'extraction dont le principe est de faire ressortir l'information pertinente par la sélection des phrases qui la caractérisent.

Les techniques d'extraction s'appuient sur :

- La combinaison des mots du titre du texte en relation avec leur présence dans le texte source [Saggion, 2000].

- L'analyse thématique du discours et de sa structure [Hernandez et Grau, 2002].
- La construction de relations de cohésion lexicale entre phrases de sorte à extraire celles qui sont le plus liées [Chali et Pinchak, 2001].
- L'utilisation de certains marqueurs représentant les relations rhétoriques comme la justification, la cause, la consécution, le contraste, la conséquence [Desclés et al., 2001],....

Ces approches peuvent être combinées en vue d'obtenir de meilleurs résumés.

Le but de ce mémoire est de mettre en oeuvre et d'évaluer un système de résumé automatique de texte en arabe en adaptant différentes techniques d'extraction qui ont déjà été utilisées en anglais.

Pour présenter nos résultats, nous procédons comme suit. Au chapitre 2, nous présentons quelques caractéristiques de la langue arabe et nous abordons la problématique du traitement morphologique qui est essentiel pour notre application. Au chapitre 3, nous introduisons les techniques d'extraction de résumé automatique intéressantes pour l'anglais et le français, et qui semblent appropriées pour la langue arabe. Le chapitre 4 décrit le système de résumé automatique de textes arabes Lakhas que nous avons développé ainsi que des résultats que nous avons obtenus. Au chapitre 5, nous décrivons les extensions à Lakhas pour participer à l'évaluation de DUC 2004 où nous avions suivi une approche originale par rapport aux autres concurrents, du fait que nous avons travaillé sur les textes dans leur langue originale, nous présentons aussi les résultats que nous avons obtenus à ce workshop. Finalement, le chapitre 6 résume les idées principales du mémoire et présente quelques axes à explorer dans les recherches futures.

2 Langue arabe

Par ses propriétés morphologiques et syntaxiques la langue arabe est considérée comme une langue difficile à maîtriser dans le domaine du traitement automatique de la langue [Aljlayl et Frieder, 2002], [Larkey et al., 2002]. L'arabe doit sa formidable expansion à partir du 7eme siècle grâce à la propagation de l'islam et la diffusion du Coran [Leclerc, 2000]. Les recherches pour le traitement automatique de l'arabe ont débuté vers les années 1970. Les premiers travaux concernaient notamment les lexiques et la morphologie.

Avec la diffusion de la langue arabe sur le Web et la disponibilité des moyens de manipulation de textes arabes, les travaux de recherche ont abordé des problématiques plus variées comme la syntaxe, la traduction automatique, l'indexation automatique des documents, la recherche d'information, etc.

> *A la différence des autres langues comme le français ou l'anglais, dont les étiquettes grammaticales proviennent d'une approche distributionnelle caractérisée par une volonté "d'écarter toute considération relative au sens", les étiquettes de l'arabe viennent d'une approche où le sémantique côtoie le formel lié à la morphologie du mot, sans référence à la position de ce dernier dans la phrase*[1].

Ce phénomène est matérialisé par les notions de schèmes et de fonctions qui occupent une place importante dans la grammaire de l'arabe.

Par exemple le mot français *ferme,* est hors contexte, un substantif, un adjectif ou un verbe. Alors que le mot arabe RaLaKa غَلَقَ est un verbe à la 3è personne masculin singulier de l'accompli actif, par contre sa forme non voyellée غلق (dans l'exemple

[1] Débili F., Achour H., Souici E. : La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique, *Correspondances de l'IRMC, N° 71, juillet-août 2002 pp.10-28.*

donné ne sont représentées que les consonnes RLK) admet quatre catégories grammaticales :

- Substantif masculin singulier (RaLKun : une fermeture),
- Verbe à la 3è personne masculin singulier de l'accompli actif (RaLaKa : il a fermé ou RaLLaKa il a fait fermé),
- Verbe à la 3è personne masculin singulier de l'accompli passif (RuLiKa : il a été fermé),
- Verbe à l'impératif 2è personne masculin singulier (RaLLiK: fais fermer).

Les voyelles jouent un rôle proche des accents en français pour un mot comme *peche* qui peut être interprété comme *pêche, pèche et péché*. Par contre, en arabe chaque lettre de chaque mot devrait posséder sa voyelle ce qui n'est en général pas le cas.

On constate donc l'étendue du rôle que jouent les voyelles dans les mots arabes, non seulement parce qu'elles enlèvent l'ambiguïté, mais aussi parce qu'elles donnent l'étiquette grammaticale d'un mot indépendamment de sa position dans la phrase.

2.1 Particularité de la langue arabe

L'alphabet de la langue arabe compte 28 consonnes (Tableau 1). L'arabe s'écrit et se lit de droite à gauche les lettres changent de forme de présentation selon leur position (au début, au milieu ou à la fin du mot). Le Tableau 2 montre les variations de la lettre ع (Ayn). Toutes les lettres se lient entre elles sauf (ا,و,ر,ز,د,ذ) qui ne se joignent pas à gauche.

Lettre arabe	Correspondant français	Prononciation
ا	a	Alef
ب	b	Ba'
ت	t	Ta'
ث	th	Tha'
ج	j	Jim
ح	h	Hha'
خ	kh	Kha'
د	d	Dal
ذ	d	Thal
ر	r	Ra
ز	z	Zayn
س	s	Sin
ش	sh	Shin
ص	s	Sad

Lettre arabe	Correspondant français	Prononciation
ض	d	Dad
ط	t	Tah
ظ	z	Zah
ع	‘ ‘	Ayn
غ	gh	Ghayn
ف	f	Fa
ق	q	Qaf
ك	k	Kaf
ل	l	Lam
م	m	Mim
ن	n	Nun
ه	h	Ha
و	w	Waw
ى	y	Ya

Tableau 1: les 28 lettres arabes. [Leclerc, 2000].

à la fin d'une lettre non joignable	à la fin	au milieu	au début
ع	ع	ـعـ	عـ

Tableau 2: Exemple de variation de la lettre ع *Ayn*

Un mot arabe s'écrit avec des consonnes et des voyelles. Les voyelles sont ajoutées au-dessus ou au-dessous des lettres (ـَ, ـِ, ـُ, ـْ). Elles sont nécessaires à la lecture et à la compréhension correcte d'un texte, elles permettent de différencier des mots ayant la même représentation. Le Tableau 3 donne un exemple pour les mots كتب et مدرسة. Cependant, les voyelles ne sont utilisées que pour des textes sacrés et didactiques. Les textes courants rencontrés dans les journaux et les livres n'en comportent habituellement pas. De plus certaines lettres comme ا Alef peuvent symboliser le أ, آ ou إ; de même que pour les lettres ى et ه qui symbolisent respectivement ي et ة [Xu et al., 2002].

Mot sans voyelles	1ère Interprétation		2ème Interprétation		3ème Interprétation	
كتب	كَتَبَ	il a écrit	كُتِبَ	Il a été écrit	كُتُبٌ	des livres
مدرسة	مَدْرَسَةٌ	école	مُدَرِّسَةٌ	enseignante	مُدَرَّسَةٌ	enseignée

Tableau 3: ambiguïté causée par l'absence de voyelles pour les mots كتب. et مدرسة

2.1.1 Morphologie arabe

Le lexique arabe comprend trois catégories de mots : verbes, noms et particules. Les verbes et noms sont le plus souvent dérivés d'une racine à trois consonnes radicales [Baloul et al., 2002]. Une famille de mots peut ainsi être générée d'un même concept sémantique à partir d'une seule racine à l'aide de différents schèmes. Ce phénomène est caractéristique à la morphologie arabe. On dit donc que l'arabe est une langue à racines réelles à partir desquelles on déduit le lexique arabe selon des schèmes qui sont des adjonctions et des manipulations de la racine. Le Tableau 4 donne quelques exemples de schèmes appliqués aux mots كتب *écrire* et حمل *porter*. On peut ainsi dériver un grand nombre de noms, de formes et de temps verbaux.

schèmes	KTB	كَتَبَ	notion d'écrire
R_1â-R_2i-R_3	KâTiB	كَاتِب	écrivain
R_1a-R_2a-R_3a	KaTaBa	كَتَبَ	a écrit
maR_1R_2aR_3	maKTaB	مَكتَب	bureau
R_1uR_2iR_3a	KuTiBa	كُتِبَ	a été écrit
…			

HML	حمل	Notion de porter
HâMiL	حَامِل	porteur
HaMaLa	حَمَلَ	a porté
maHMaL	مَحمَل	brancard
HuMiLa	حُمِلَ	a été porté

Tableau 4: Exemple de schèmes pour les mots كتب *écrire* et حمل *porter*

Les lettres en majuscule (R_i) désignent les consonnes de base qui composent la racine.

Les voyelles (â, a, i,..) désignent les voyelles et les consonnes en minuscule (m,..) sont des consonnes de dérivation utilisées dans les schèmes.

La majorité des verbes arabes ont une racine composée de 3 consonnes. L'arabe comprend environ 150 schèmes ou patrons dont certains plus complexes, tel le redoublement d'une consonne ou l'allongement d'une voyelle de la racine,

l'adjonction d'un ou de plusieurs éléments ou la combinaison des deux. Une autre caractéristique est le caractère flexionnel des mots : les terminaisons permettent de distinguer le mode des verbes et la fonction des noms [Baloul et al., 2002].

2.1.2 Structure d'un mot

En arabe un mot peut signifier toute une phrase grâce à sa structure composée qui est une agglutination d'éléments de la grammaire, la représentation suivante schématise une structure possible d'un mot. Notons que la lecture et l'écriture d'un mot se fait de droite vers la gauche.

Post fixe	Suffixe	Corps schématique	Préfixe	Antéfixe

- Antéfixes sont des prépositions ou des conjonctions.
- Préfixes et suffixes expriment les traits grammaticaux et indiquent les fonctions : cas du nom, mode du verbe et les modalités (nombre, genre, personne,…)
- Postfixes sont des pronoms personnels.

Exemple

أَتَتَذَكَّرُونَنَا

Ce mot exprime la phrase en français : "*Est ce que vous vous souvenez de nous ?*"

La segmentation de ce mot donne les constituants suivants :

أ | دَ | تَذَكَّرُ | ونَ | نَا

Antéfixe : أ conjonction d'interrogation

Préfixe : دَ préfixe verbal du temps de l'inaccompli.

Corps schématique: تَذَكَّر dérivé de la racine: ذكر selon le schème $taR_1aR_2aR_3a$

Suffixe : ون suffixe verbal exprimant le pluriel

Post fixe : نا pronom suffixe complément du nom

2.1.3 Catégories des mots

L'arabe considère 3 catégories de mots

- Le verbe : entité exprimant un sens dépendant du temps, c'est un élément fondamental auquel se rattachent directement ou indirectement les divers mots qui constituent l'ensemble.
- Le nom : l'élément désignant un être ou un objet qui exprime un sens indépendant du temps.
- Les particules : entités qui servent à situer les événements et les objets par rapport au temps et l'espace, et permettent un enchaînement cohérent du texte.

2.1.3.1 Le verbe

La plupart des mots en arabe, dérivent d'un verbe de trois lettres. Chaque verbe est donc la racine d'une famille de mots. Comme en français, le mot en arabe se déduit de la racine en rajoutant des suffixes ou des préfixes.

La conjugaison des verbes dépend de plusieurs facteurs :

- Le temps (accompli, inaccompli).
- Le nombre du sujet (singulier, duel, pluriel).
- Le genre du sujet (masculin, féminin).
- La personne (première, deuxième et troisième)
- Le mode (actif, passif).

Par exemple : ب + ت + ك *K+T+B* donne le verbe كتب *KaTaBa.* (écrire).

Dans tous les mots qui dérivent de cette racine, on trouvera ces trois lettres K, T, B (voir Tableau 4).

La conjugaison des verbes se fait en ajoutant des préfixes et des suffixes, un peu comme en français.

La langue arabe dispose de trois temps.

- L'accompli : correspond au passé et se distingue par des suffixes (par exemple pour le pluriel féminin on a كتبن KaTaBna, *elles ont écrit* et pour le pluriel masculin on a كتبوا KaTaBuu, *ils ont écrit*).
- L'inaccompli présent: présente l'action en cours d'accomplissement, ses éléments sont préfixés (يكتب yaKTuBu *il écrit*; تكتب taKTuBu, *elle écrit*).
- L'inaccompli futur : correspond à une action qui se déroulera au futur et est marqué par l'antéposition de س *sa* ou سوف *sawfa* au verbe (سيكتب sayaKTuBu *il écrira*, سوف يكتب sawfa yaKTuBu *il va écrire*).

2.1.3.2 Les noms

Les substantifs arabes sont de deux catégories, ceux qui sont dérivés de la racine verbale et ceux qui ne le sont pas comme les noms propres et les noms communs. Dans le premier cas, le fait que le nom soit dérivé d'un verbe, il exprime donc une certaine sémantique qui pourrait avoir une influence dans la sélection des phrases saillantes d'un texte pour le résumé.

La déclinaison des noms se fait selon les règles suivantes:

- Le féminin singulier: On ajoute le ة, exemple صغير *petit* devient صغيرة *petite*
- Le féminin pluriel : De la même manière, on rajoute pour le pluriel les deux lettres ات, exemple صغير *petit* devient صغيرات *petites*
- Le masculin pluriel : Pour le pluriel masculin on rajoute les deux lettres ين ou ون dépendamment de la position du mot dans la phrase (sujet ou complément d'objet), exemple : الراجع *revenant* devient الراجعين ou الراجعون *revenants*
- Le Pluriel irrégulier: Il suit une diversité de règles complexes et dépend du nom. exemple :طفل *un enfant* devient أطفال *des enfants*

Le phénomène du pluriel irrégulier dans l'arabe pose un défi à la morphologie, non seulement à cause de sa nature non concaténative, mais aussi parce que son analyse dépend fortement de la structure [Kiraz, 1996] comme pour les verbes irréguliers.

Certain dérivés nominaux associent une fonction au nom :

- Agent (celui qui fait l'action),
- Objet (celui qui a subit l'action),
- Instrument (désignant l'instrument de l'action),
- Lieu.

Pour les pronoms personnels, le sujet est inclus dans le verbe conjugué. Il n'est donc pas nécessaire (comme c'est le cas en français) de précéder le verbe conjugué par son pronom. On distinguera entre singulier, duel (deux) et pluriel (plus de deux) ainsi qu'entre le masculin et féminin.

2.1.3.3 Les particules

Ce sont principalement les mots outils comme les conjonctions de coordination et de subordination.

Les particules sont classées selon leur sémantique et leur fonction dans la phrase, on en distingue plusieurs types (introduction, explication, conséquence, ...). Elles jouent un rôle important dans l'interprétation de la phrase [Kadri et Benyamina, 1992]. Elles servent à situer des faits ou des objets par rapport au temps ou au lieu, elles jouent également un rôle clé dans la cohérence et l'enchaînement d'un texte.

Comme exemple de particules qui désignent un temps منذ , قبل , بعد *pendant, avant, après*, un lieu حيث *où*, ou de référence الذين *ceux*,....

Ces particules seront très utiles pour notre traitement à deux niveaux :

- Elles font partie de l'antidictionnaire qui regroupe les termes à ne pas prendre en considération lors de calcul de fréquence de distribution des mots,
- Elles identifient des propositions composant une phrase.

Les particules peuvent avoir des préfixes et suffixes ce qui rajoute une complexité quant à leur identification.

2.2 Problèmes du traitement automatique de l'arabe

Un des aspects complexes de la langue arabe est l'absence des voyelles dans le texte, qui risque de générer une certaine ambiguïté à deux niveaux :

- Sens du mot
- Difficulté à identifier sa fonction dans la phrase, (différencier entre le sujet et le complément,...).

Ceci peut influencer les fréquences des mots étant donné qu'elles sont calculées après la détection de la racine ou la lemmatisation des mots qui est basée sur la suppression de préfixes et suffixes. Lors du calcul des scores à partir des titres, il peut arriver que des mots soient considérés comme dérivant d'un même concept alors qu'ils ne le sont pas.

Dans l'exemple 1, en utilisant la distribution des mots ou le titre avec ou sans lemmatisation, la phrase 3 aura un score le plus important alors que les phrases 1 et 2 semblent plus intéressantes, ce qui n'aurait pas été le cas avec un texte voyellé.

العنوان : اثر العلم 1- العلماء..... 2- علميا..... 3- في المحاضرة ليس العلم الوطني ولكن العلم لكل الدول ....-	Titre : impact de la science 1 - Les scientifiques .... 2 - Scientifiquement.... 3 - A la conférence non seulement le drapeau national... mais aussi le drapeau de chaque pays.... - ...

Exemple 1: Effet du mot non voyellé العلم sur les extraits.

L'ambiguïté vient du mot العلم *la science* ou *drapeau* alors que voyellé on aura العِلْمُ pour *la science* et العَلَمُ pour *le drapeau*.

Cette ambiguïté pourrait, dans certains cas, être levée soit par une analyse plus profonde de la phrase ou des statistiques (par exemple il est plus probable d'avoir العلم الوطني *le drapeau national* que la *science nationale*).

De plus la capitalisation n'est pas employée dans l'arabe ce qui rend l'identification des noms propres, des acronymes, et des abréviations encore plus difficile [Hammou et al., 2002].

Comme la ponctuation est rarement utilisée, on doit ajouter une phase de segmentation de phrase pour l'analyse d'un texte [Chalabi, 2001].

2.2.1 Segmentation de phrase

La reconnaissance de la fin de phrase est délicate car la ponctuation n'est pas systématique et parfois les particules délimitent les phrases.

Pour la segmentation de texte [Ouersighni, 2001] utilise :

- Une segmentation morphologique basée sur la ponctuation,
- Une segmentation basée sur la reconnaissance de marqueurs morphosyntaxiques ou des mots fonctionnels comme : أو , و , أي , لكن , حتى *ou, et, c.a.d., mais, quand.*

Cependant, ces particules peuvent jouer un autre rôle que celui de séparer les phrases.

2.2.2 Détection de racine

Pour détecter la racine d'un mot, il faut connaître le schème par lequel il a été dérivé et supprimer les éléments flexionnels (antéfixes, préfixes, suffixes, post fixes) qui ont été ajoutés.

Nous utilisons la liste de préfixes et de suffixes proposé par [Darwish, 2003] voir Tableau 5. Plusieurs d'entre eux ont été utilisés par [Chen et Gey, 2002] pour la

lemmatisation de mots arabes; ils ont été déterminés par un calcul de fréquence sur une collection d'articles arabes de l'Agence France Press (AFP).

Préfixes							
لا	في	لل	كم	بم	وت	بت	وال
با	وا	ليـ	فم	لم	ست	يت	فال
	فا	ويـ	ال	وم	نت	مت	بال
Suffixes							
ا	ة	ين	ية	هم	ته	وه	ات
	ه	يه	تك	هن	تم	ان	وا
	ي	ية	نا	ها	كم	تي	ون

Tableau 5: Liste des préfixes et suffixes les plus fréquents

L'analyse morphologique devra donc séparer et identifier des morphèmes semblables aux mots préfixés comme les conjonctions wa-و et fa-ف, des prépositions préfixées comme bi-ب et li-ل, l'article défini ال, des suffixes de pronom possessif.

La phase d'analyse morphologique détermine un schème 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 [Darwish, 2002].

Lorsqu'un mot peut être dérivé de plusieurs racines différentes, la détection de la racine est encore plus difficile, en particulier en absence de voyelles [Attia, 2000].

Par exemple, pour le mot arabe ايمان AymAn les préfixes possibles sont : "∅","A ا" et "Ayاي" et les suffixes possibles sont : "∅" et "Anان" (Tableau 6), sans compter que ce mot peut aussi représenter un nom propreإيمَان *Imène*.

Stem	Préfixe	Schème	Suffixe	Racine	signification
AymAn ايمان	∅	R1yR2aR3	∅	Amn امن	croyance
ymAn يمان	ا A	R1R2aR3	∅	Ymn يمن	convenant
mAn مان	اي Ay	R1R2R3	∅	mAn مان	Va t il approvisionner
Aym ايم	∅	R1R2R3	ان An	Aym ايم	Deux veuves

Tableau 6: Les stems possibles pour le mot ايمان .

Certains verbes sont considérés comme irréguliers, ce sont ceux qui portent des consonnes particulières dites faibles (و, ا, ي). Ils sont appelés ainsi parce que, lors de leur déclinaison, chacune de ces lettres est soit conservée, soit remplacée ou éliminée [Kadri et Benyamina, 1992]. Le Tableau 7 donne un exemple de dérivation du mot قال *dire*.

Caractère ا est remplacé par	قال	*dire*
ا	قال	*il a dit*
و	يقول	*il dit*
ي	قيل	*il a été dit*
∅	قل	*dis*

Tableau 7: Exemple de déclinaison du verbe irrégulier قال *dire*

Une difficulté en traitement automatique de l'arabe est l'agglutination par laquelle les composantes du mot sont liées les unes aux autres. Ce qui complique la tache de l'analyse morphosyntaxique pour identifier les vrais composants du mot.

Par exemple, le mot ٱلَمُهُمْ ALaMuhum, *leur douleur* dans sa forme voyellée n'accepte qu'une seule segmentation : هُمْ + ٱلَمُ (ALaMu+hum).

Dans sa forme non voyellée المهم (ALMHM), le même mot accepte au moins les trois segmentations présentées dans le Tableau 8.

Segmentation possible		Traduction en français
أ + لم + هم	a+LM+hm	*les a-t-il ramassés*
ألم + هم	ALM+hm ALM+hm	*leur douleur* *il les a fait souffrir*
أل + مهم	al+MHM	*l'important*

Tableau 8: Exemple de segmentation du mot المهم

L'amplification de l'ambiguïté de segmentation s'opère selon deux façons [Débili et al., 2002]:

- d'abord, il y a plus d'unités ambiguës dans un texte non voyellé que dans son correspondant voyellé,
- mais aussi, les unités ambiguës acceptent plus de segmentations dans le texte non voyellé.

De plus le fait de précéder la lemmatisation par la troncature des préfixes avant les suffixes (et réciproquement) peut influencer les résultats. En considérant l'exemple dans le Tableau 8, sur un texte où la notion de douleur est importante, le fait d'avancer la suppression des préfixes avant les suffixes les mots comme المهم *leur douleur (pour le pluriel)*, المهما *leur douleur* (pour le duel) exprimeront une toute autre notion.

2.3 Conclusion

Plusieurs travaux sur la langue arabe ont été développés récemment dans le domaine de la traduction automatique et de la recherche d'information ; ce qui a en fait ressortir la difficulté à cause de l'ambiguïté due à l'absence de voyelles amplifiée par l'agglutination des mots par rapport à d'autres langues comme le français ou l'anglais.

Bien que la lemmatisation soit difficile pour les langues avec des morphologies complexes comme l'arabe, elle est particulièrement importante et utile en particulier dans les systèmes de recherche d'information. Il est suffisant de regrouper les mots qui se ressemblent le plus sans pour autant connaître la racine exacte.

[Larkey et al., 2002] suggèrent que l'utilisation de stop-words donne de meilleurs résultats sur des mots arabes lemmatisés que sur des mots non lemmatisés.

Contrairement à l'anglais, la langue arabe possède un système dérivationnel très riche, et c'est dans cette caractéristique que réside la difficulté de traiter cette dernière.

En ce qui concerne les études sur les résumés automatiques sur la langue arabe, elles n'en sont qu'à leur début. Il n'y a pas encore des travaux avancés qui exploitent ou adaptent les techniques de résumés automatiques actuelles, encore moins en ce qui concerne la génération et la condensation. Néanmoins pour cette année, il y a eu plusieurs travaux qui ont été présentés dans des conférences et workshops qui ont été organisées spécialement pour le traitement de la langue arabe (*JEP-TALN 2004: Le traitement automatique de la langue arabe à Fès Maroc, COLING 2004: Computational Approaches to Arabic Script-based Languages à Genève Suisse, Arabic Language Resources and Tools Conference au Caire Égypte*). Ce qui va permettre de rattraper le retard en développant des techniques plus puissantes tout en offrant des évaluations plus formelles.

3 Résumé automatique

Le but d'un résumé automatique de texte est de produire une représentation abrégée d'un ou de plusieurs documents.

Les premiers travaux sur les résumés automatiques de textes datent des années cinquante. Pour extraire les phrases pertinentes nécessaires à la construction d'un résumé, [Luhn, 1958] considère des caractéristiques comme la fréquence d'occurrence des termes, des mots de titres et la position de la phrase.

Avec l'avènement de l'Internet et de moteurs de recherche de plus en plus performants, l'importance d'informations condensées du type résumé est devenue nécessaire pour faire ressortir l'information pertinente. De ce fait le résumé automatique a inspiré de nouvelles orientations, plusieurs nouvelles approches ont commencé à être explorées en linguistique (basée sur l'analyse du discours et de sa structure) et en statistique (basée sur la distribution des occurrences des mots) [Amini et Gallinari, 2002].

Le résumé de texte automatique peut être classifié en deux approches : abstraction et extraction.

Dans l'abstraction, le texte résumé est une interprétation du texte original avec un processus de production par récriture du texte source en une version plus courte par le remplacement de certains concepts. Sa mise en œuvre exige l'utilisation de grammaires et de lexiques pour l'analyse syntaxique et la génération, en plus d'une modélisation de la compréhension humaine des textes. Ce processus est très difficile à mettre en œuvre.

L'extraction est le processus consistant à choisir des extraits appropriés (des phrases, des paragraphes, etc.) du texte original et à les enchaîner dans une forme plus courte [Jaruskulchai et Kruengkrai, 2003]. Le texte résumé est extrait du texte sur une base

statistique ou en employant des méthodes heuristiques ou une combinaison des deux. Souvent, on extrait du texte source les phrases complètes jugées les plus importantes. Cette approche a l'avantage d'être facile à réaliser mais elle risque d'introduire une certaine incohérence dans les résumés.

La plupart des travaux récents dans ce secteur de recherche sont basés sur l'extraction. Bien que la lecture des résumés par extraction soit difficile en raison du manque de cohérence, elle dépend aussi de l'objectif du résumé : indicatif ou informatif.

Le résumé indicatif décrit le contenu du texte et aide le lecteur à décider s'il doit consulter le document original ou pas. Il lui fournit une idée du texte sans donner le contenu spécifique et signale les thèmes du document dans un style parfois télégraphique. C'est une sorte d'indexe au texte source, c'est-à-dire qu'il transmet seulement les sujets principaux.

Le résumé informatif cherche à condenser le texte de façon à ce que le lecteur n'ait pas besoin d'aller consulter le document original [Torres et al., 2001]. Les informations sont présentées dans l'ordre du document, mais leur importance relative peut différer de celle du document.

3.1 Les méthodes de résumé

Dans cette partie, nous présentons brièvement différentes méthodes employées pour l'extraction de phrases clefs; elles sont basées essentiellement sur le calcul d'un score associé à chaque phrase afin d'estimer son importance dans le texte. Le résumé final ne gardera que les phrases avec les meilleurs scores.

3.1.1 Méthodes à base de mots clés

Cette méthode est basée sur le fait que l'auteur se sert (pour exprimer ses idées principales) de quelques mots-clés qui ont tendance à être récurrents dans le texte [Pardo et al., 2002]. Le résumé automatique est alors produit en recherchant dans le

texte source les unités de texte minimales réunissant ses mots-clés. Ce principe est souvent appliqué en différentes variantes présentées dans les sous-sections qui suivent.

3.1.1.1 Mots-clé prédéfinis

Pour calculer le score de chaque phrase S selon les mots-clés qu'elle contient, on peut calculer le score suivant :

$$\text{Score}_{\text{mot-cle}}(S) = \sum_{w \in S} a(w) \times F(w)$$

$$a(w) = \begin{cases} A & \text{si } w \in \text{liste de mots - clés } (A > 1) \\ 1 & \text{sinon} \end{cases}$$

$F(w)$ est la fréquence du terme w dans la phrase S

La liste de mots-clés peut être introduite par l'utilisateur (domaine d'intérêt) ou composée des mots-clés établis par l'auteur. L'importance du poids du terme w est donné par $A \times F(w)$, avec $A > 1$.

3.1.1.2 Titres

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. Par conséquent, on peut attribuer à chaque phrase un poids en fonction de sa ressemblance avec le titre [Ishikawa et al., 2001].

Dans ce cas on considère les mots du titre du texte comme des mots-clés et on produit le résumé en sélectionnant les phrases qui couvrent certains mots apparaissant dans un titre.

$$\text{Score}_{\text{titre}}(S) = \sum_{w \in S} b(w) \times F(w)$$

$$b(w) = \begin{cases} A & \text{si } w \in \text{liste de mots du titre } (A > 1) \\ 1 & \text{sinon} \end{cases}$$

3.1.1.3 Distribution des termes

L'idée de cette méthode est de considérer comme importantes *les phrases qui contiennent des mots importants* du texte. Un mot est considéré important s'il est employé assez fréquemment dans le texte.

Pour le calcul des fréquences on considère généralement les mots qui appartiennent à des classes non fermées de la langue telles que les noms et les verbes [Saggion, 2000].

$$\text{Score}_{tf.idf}(S) = \frac{1}{|S|} \sum_{w \in S} tf.idf(w)$$

$$tf.idf(w) = \frac{tf(w)-1}{tf(w)} \log \frac{DN}{df(w)}$$

$|S|$ = nombre de mots dans la phrase S

$tf(w)$ est la fréquence du terme w dans le document

$df(w)$ est le nombre de documents du corpus où le terme w apparaît

DN est le nombre de documents dans le corpus

Il y a plusieurs fonctions de calcul de *tf.idf*, pour Nobata*y* et Sekine [Nobata et Sekine, 2003] cette dernière semble donner de meilleures performances sur leur corpus (journal de Wall Street de 1994 à 1995).

Au lieu de considérer un document par rapport à une collection, une autre méthode considère une phrase par rapport à un texte [Pardo et al., 2002].

$$\text{Score (S)} = \frac{1}{|S|}\sum_{w \in S} \text{Score}(w)$$

$$\text{Score}(w) = \text{F}(w) \times \frac{\log(|S|)}{\text{S}(w)}$$

F(w) est la fréquence du terme w dans la phrase

S(w) nombre de phrases dans lesquelles w apparaît .

3.1.2 Méthode à base de position

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 ; 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 [Ishikawa et al., 2001].

On définit le score d'une phrase S à la position i comme suit :

$$\text{Score}_{\text{lead}}(S_i) = \beta_i$$

$$\beta_i = \begin{cases} B \succ 0 & \text{si } i \prec N \\ 0 & \text{si } i \geq N \end{cases}$$

β_i est une fonction rectangulaire qui modélise la distribution de phrases importantes selon leur position dans l'article.

Dans le cas où les dernières phrases auraient une certaine importance, il suffit d'introduire un nouvel intervalle pour la valeur de i.

L'inconvénient de cette méthode est qu'elle dépend de la nature du texte à résumer ainsi que du style de l'auteur.

3.1.3 Méthode dépendant de la longueur de phrase

Cette méthode attribue un poids à une phrase en fonction du nombre de mots dans la phrase [Nobata et Sekine, 2003].

Deux techniques peuvent être employées pour le calcul du score :

- longueur de chaque phrase (L_i) par rapport à la longueur maximale de la phrase. $\text{Score}_{\text{long}}(S_i) = L_i/L_{\max}$
- affecte un score nul à une phrase plus courte qu'une certaine longueur (L minimale) :

$$\text{Score}_{\text{long}}(S_i) = \begin{cases} 0 & \text{si } L_i \leq L_{\min} \\ \dfrac{L_i - L_{\min}}{L_{\min}} & \text{si } L_i \succ L_{\min} \end{cases}$$

3.1.4 Méthode à base d'expressions indicatives (cue methods)

Cette méthode choisit des unités de texte avec des indications spécifiques ou des expressions spécifiques. Par exemple, pour les textes scientifiques, on a comme expressions *le but de ce travail ...*, *ce papier présente ...*, *les résultats* et *des conclusions* sont de bons candidats pour indiquer les phrases à inclure dans un résumé. Des textes de types différents peuvent avoir des expressions indicatives différentes.

On peut déduire un score pour une phrase d'un texte quelconque à analyser en fonction de la ressemblance qu'elle présente, pour le trait donné.

On pourrait définir le score d'une phrase S correspondant à un certain motif comme:

$$\text{Score}_{cue}(S) = \begin{cases} 1 & \text{si } S \text{ correspond à un motif} \\ 0 & \text{sinon} \end{cases}$$

3.1.5 Méthode basée sur les relations (cohésion lexicale)

L'exploitation des fréquences de mots est un bon moyen pour faire ressortir les termes clés dans un texte mais elle ne prend pas en compte les relations entre les mots dans les différentes parties du texte. L'extraction de phrases basée sur la fréquence de mots cause souvent un manque de cohésion. Pour pallier ce problème, on a développé une approche basée sur la cohésion grammaticale (c'est-à-dire, la référence, la substitution, la conjonction) et la cohésion lexicale (c'est-à-dire, des mots liés sémantiquement).

Cette méthode suggère que plus une phrase est liée à une autre dans un texte, plus elle est appropriée dans ce contexte c'est-à-dire qu'elle exprime le même sujet. Ainsi, des phrases liées doivent être choisies ensemble pour composer un résumé. L'omission de certaines phrases fortement corrélées pourrait produire des textes incohérents. L'identification de telles corrélations est basée normalement sur un thésaurus ou lexique informatisé qui permet de déterminer les relations entre les mots. On construit des chaînes lexicales à partir des mots candidats du texte, ces chaînes regroupent des mots liés par des relations obtenues à partir du thésaurus. Les phrases qui sont le plus connectées aux chaînes lexicales sont extraites [Chali et Pinchak, 2001], [Pardo et al., 2002].

3.1.6 La méthode d'exploration contextuelle

La méthode d'exploration contextuelle vise à identifier les connaissances linguistiques dans le texte en les restituant dans leurs contextes et en les organisant en tâches spécialisées. L'approche est fondée sur la construction manuelle d'une base de données de marqueurs linguistiques et une expression de règles d'exploration contextuelle. Ces règles appliquées aux phrases du texte source vont filtrer les informations sémantiques indépendantes du domaine avec les étiquettes

sémantiques hiérarchisées comme : énoncés structurants, définition, causalité, etc. La stratégie de sélection des unités saillantes est fonction des besoins des utilisateurs [Farzindar, 2003].

L'exploration contextuelle appuie son analyse sur une hiérarchisation des connaissances. À cet effet, on distingue quatre niveaux de connaissances [Berri, 1996]:

- les connaissances linguistiques (grammaticales et lexicales), indépendantes des connaissances sur le monde externe;
- les connaissances propres à un domaine particulier: elles concernent le savoir-faire lié au domaine de compétence et les règles qui organisent ce domaine;
- les connaissances socioculturelles qui dépendent de l'environnement social, des usages, des coutumes, etc.;
- les connaissances encyclopédiques qui sont générales et communes à tous les êtres humains.

Les stratégies décisionnelles de l'exploration contextuelle s'expriment sous la forme de règles heuristiques qui identifient en premier lieu un indicateur pertinent, caractéristique du problème à résoudre. Ensuite le contexte linguistique est fouillé pour rechercher des indices linguistiques afin de prendre une décision adéquate [Berri, 1996].

Les avantages de cette technique sont : l'indépendance entre les connaissances linguistiques nécessaires au système et les connaissances accumulées sur un domaine particulier. Elle permet une extensibilité incrémentale, en complétant les listes déjà établies (recherche d'indices plus fins) et en affinant les règles d'exploration (Desclés). Un Système d'Exploration Contextuelle est donc plus ou moins performant selon la richesse des indices pris en compte et la finesse de l'exploration.

3.1.7 Méthode hybride

Les méthodes présentées dans les sections précédentes utilisent des traits (fréquence, position, expression indicative, etc.) qui ne peuvent isolément garantir des résultats optimaux [Saggion, 2000]. On combine souvent ces traits par exemple avec l'équation suivante :

$$\text{Score}_{\text{hybride}}(S) = a_1 * \text{Score}_{\text{tf*idf}}(S) + a_2 * \text{Score}_{\text{lead}}(S) + a_3 * \text{Score}_{\text{cue}}(S) + a_4 * \text{Score}_{\text{titre}}(S)$$

Équation 1: Fonction de calcul du score global

Les poids a_i peuvent être fixés arbitrairement ou déterminés de manière expérimentale (par apprentissage par exemple).

Certaines expériences de [Edmundson, 1969] sur un corpus hétérogène de 200 documents ont montré que si on combine les méthodes cue, titre et position (poids zéro pour la méthode mot-clés), on obtient de meilleurs résultats que si on les combine avec la méthode mot-clés.

Dans le cas de textes journalistiques, [Strzalkowski et al., 1998] ont combiné les méthodes de distribution de termes, du titre, de la position et de la cue, en considérant la spécificité du texte. Ils ont fait ressortir que les phrases qui commencent par des nominaux ou contiennent des mots du titre semblent être plus pertinentes que des phrases n'ayant pas ce caractère. De plus, les mots ou les phrases n'apparaissant que dans quelques paragraphes sont plus importants que ceux mentionnés dans tous les paragraphes. Pour garder la cohérence du texte, le résumé est composé d'une sélection de paragraphes pertinents.

3.2 Conclusion

Nous avons présenté quelques techniques pour le résumé automatique de textes : les méthodes d'extraction offrent certains avantages: simplicité de mise en œuvre, rapidité de traitement, indépendance des traitements par rapport à la langue et une compression paramétrable en modifiant les seuils de sélection. Certaines méthodes

semblent offrir de meilleurs résultats que d'autres, cela est dû en grande partie à la nature du texte et au style de l'auteur.

L'exploration contextuelle semble assez puissante mais demande une analyse profonde de la langue et une base de connaissance appropriée aux types de texte. L'approche mixte (hybride) est souvent utilisée : comme les coefficients des paramètres sont déterminés expérimentalement, on arrive à obtenir de bons résultats. C'est pourquoi nous avons choisi cette méthode pour notre travail sur les résumés automatique de textes arabes.

Notre méthode est basée essentiellement sur les méthodes d'extractions, qui ont prouvé leur efficacité pour d'autres langues et semblent donner des résultats satisfaisants au niveau informationnel.

4 Architecture globale de Lakhas

Lakhas est un système de résumé automatique de texte en langue Arabe basé principalement sur des techniques d'extraction. La mise en œuvre fonctionnelle de LAKHAS est représentée à la figure 1. Elle repose sur une segmentation à différents niveaux ainsi que sur le calcul des poids afin de permettre la génération de résumé. LAKHAS est flexible et comporte plusieurs modules qui peuvent communiquer entre eux. Nous décrivons maintenant brièvement les modules selon la numérotation de la figure 1.

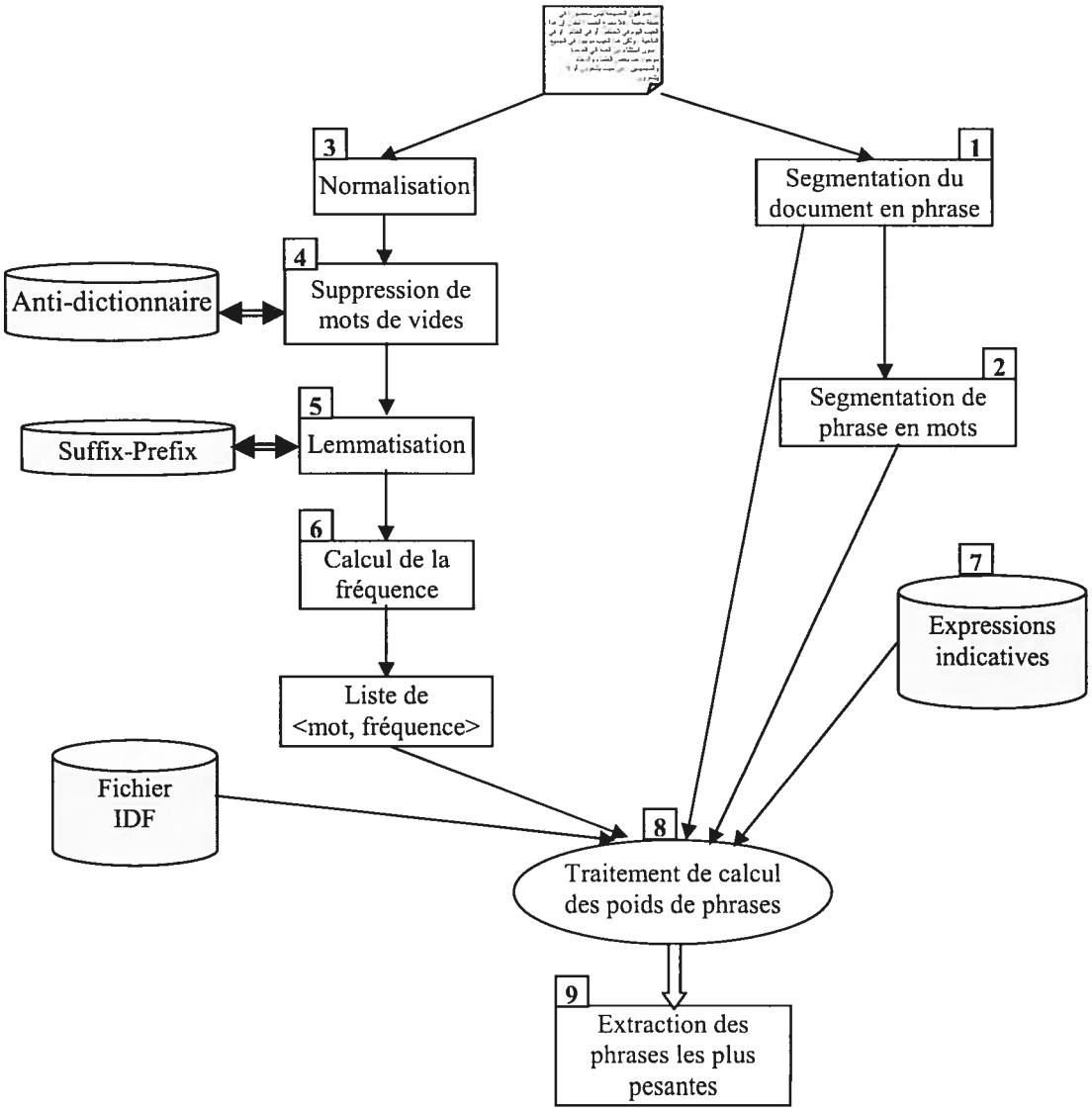

Figure 1: Schéma global de LAKHAS, les numéros des modules sont référencés par items de la section 4.1

4.1 Description des principaux modules composant Lakhas

1. **Segmentation du document en phrases** identifie chaque phrase du document dans le corpus grâce aux balises <P> et </P>.

2. **Segmentation des phrases en mots,** appelée aussi *tokenisation* sépare chaque phrase en une séquence de mots, en détectant les délimiteurs de mots tels que l'espace ou la ponctuation. Ceci permet de retourner une liste de mots avec leur fréquence par rapport à la phrase.

3. **Normalisation** transforme une copie du document original dans un format standard plus facilement manipulable. Avant la lemmatisation, le document est normalisé comme suit [Larkey et al., 2002]:
    - Suppression des caractères spéciaux et les chiffres
    - Remplacement de ٱ, إ et أ avec ا
    - Remplacement de la lettre finale ي avec ى
    - Remplacement de la lettre finale ة avec ه

    Cette étape est nécessaire à cause des variations qui peuvent exister lors de l'écriture d'un même mot arabe (section 2.1). L'extraction se fait à partir du document original ce qui permet de préserver l'intégralité de l'information.

4. **Suppression des mots vides** consiste à éliminer tous les mots non significatifs. Pour chaque mot reconnu, on le compare avec un des éléments dans l'antidictionnaire qui contient tous les mots non-significatifs. Si un mot en fait partie, il ne sera pas pris en considération pour le calcul de sa fréquence. L'antidictionnaire regroupe en particulier les particules (section 2.1.3.3) pour lesquelles nous avons ajouté quelques flexions possibles.

5. **Lemmatisation** Pour un mot significatif, on applique une lemmatisation légère qui consiste à essayer de déceler si des préfixes ou suffixes ont été ajoutés au mot [Darwish, 2002]. Puisque la plupart des mots arabes ont une

racine à trois ou quatre lettres, le fait de garder le mot au minimum à trois lettres va permettre de préserver l'intégrité du sens du mot.

La liste que nous utilisons (section 2.2.2) regroupe les préfixes et les suffixes les plus utilisés dans la langue arabe tels que les conjonctions, préfixes verbaux, pronoms possessifs, pronoms compléments du nom ou suffixes verbaux exprimant le pluriel etc…

6. **Calcul de fréquence** calcule le nombre d'occurrences d'un mot significatif dans un document.

7. **Expressions indicatives** augmentent le poids des phrases qui pourraient apporter une information intéressante. C'est en quelque sorte le rôle inverse des mots vides. On utilise un dictionnaire contenant l'ensemble de ces expressions. Il regroupe les annonces thématiques, les soulignements etc.

   Vu la particularité de la langue arabe au niveau morphologique (agglutination des mots), nous avons enrichi ce dictionnaire de variantes des expressions, en arabe les articles, les pronoms, les conjonctions et les particules se collent au mot comme le montre le Tableau 9.

Variante de اشار	Signification en français
اشار	*il a signalé*
واشار	*et il a signalé*
اشارت	*elle a signalé*
فاشار	*alors il a signalé*

Tableau 9: les variations possibles du mot اشار *signaler*

8. **Calcul du poids des phrases**

   Le Score global sera calculé par combinaison des différentes méthodes grâce à l'équation 1 montrée dans la section 3.1.7.

Les coefficients a_i sont initialement fixés à 1. L'avantage de cette technique est qu'elle permet d'ajuster ces coefficients suivant la nature des corpus. Le résultat sera retourné sous forme de liste de phrases triée par score.

9. **Extraction des phrases** permet de retourner le résultat final suivant le choix du pourcentage de compression. Ce pourcentage représente le nombre de phrases extraites par rapport au nombre de phrases contenues dans le document.

4.2 Corpus d'application AFP (Agence France Presse)

Le Corpus arabe AFP est l'une des archives des données de texte de type *newswire* qui ont été acquises par le Consortium de Données Linguistique (LDC) à partir des sources de nouvelles arabes. Nous avons sélectionné l'ensemble des documents de l'année 2000 soit 54 726 fichiers, sur lesquelles nous avons effectué les calculs du fichier IDF (fréquence des documents par rapport à un terme). La Figure 2 montre un document du corpus visualisé par le navigateur Web de Microsoft.

Figure 2: Fichier visualisé par un browser

Toutes les données de texte sont représentées dans le format SGML avec un encodage UTF-8. La Figure 3 représente le fichier source avec ses balises.

```
<DOC id="AFA19981111.0059" type="story" >
<HEADLINE>
Ø¨Ø±Ø§ÙƒØ§Ø´ Ø´Ø§Ù‡: Ø¨Ø²Ø¯Ø§Ø¯ ØªØ±Ù Ø¶ Ø§Ù„Ø¹Ø±Ø§Ø¬Ø¹
</HEADLINE>
<DATELINE>
Ø¨Ø²Ø¯Ø§Ø¯ 11-11 (Ø§Ù  Ø¨)   [Ø¹Ø±Ø§ NXT38 Ø§Ù„Ø¹Ø±Ø§Ù,/Ø§Ù„Ø§Ù…Ù…-Ø§Ù„Ù…Ø¹ØØ¯Ø© Ù…Ù^Ø³Ø¹-
Ø«Ø§Ù†]
</DATELINE>
<TEXT>
<P>
Ø§Ø¹Ù„Ù† Ø§Ù„Ù…Ù…Ø«Ù„ Ø§Ù„Ø®Ø§Øµ Ù„Ù„Ø§Ù…ÙŠÙ† Ø§Ù„Ø¹Ø§Ù… Ù„Ù„Ø§Ù…Ù… Ø§Ù„Ù…Ø¹ØØ¯Ø© Ù ÙŠ
Ø¨Ø²Ø¯Ø§Ø¯
Ø¨Ø±Ø§ÙƒØ§Ø´ Ø´Ø§Ù‡ Ø§Ù„ÙŠÙˆÙ… Ø§Ù„Ø§Ø±Ø¨Ø¹Ø§Ø¡ Ø¨Ø¹Ø¯ Ù„Ù‚Ø§Ø¡ Ù…Ø¹ Ù†Ø§Ø Ø¨ Ø±Ø ÙŠØ³
Ø§Ù„Ù^Ø²Ø±Ø§Ø¡ Ø§Ù„Ø¹Ø±Ø§Ù‚ÙŠ Ø·Ø§Ø±Ù‚ Ø¹Ø²ÙŠØ² Ø§Ù†
Ø§Ù„Ø¹Ø±Ø§Ù‚ ÙŠØ±Ù Ø¶ Ø§Ù„Ø¹Ø±Ø§Ø¬Ø¹ Ø¹Ù† Ù‚Ø±Ø§Ø±Ù‡ Ø¨Ù‚Ø·Ø¹ Ø§Ù„Ø¹Ø¹Ø§Ù^Ù† Ù…Ø¹ Ù…Ù Ø¹Ø´ÙŠ
Ù†Ø²Ø¹ Ø§Ù„Ø§Ø³Ù„ØØ© Ø§Ù„Ø¯Ù^Ù„ÙŠÙŠÙ†.
</P>
<P>
Ù^Ù‚Ø§Ù„ Ø´Ø§Ù‡ Ø¨Ø¹Ø¯ Ù„Ù‚Ø§Ø¡ Ù…Ø¹ Ø·Ø§Ø±Ù‚ Ø¹Ø²ÙŠØ² "Ù„Ù‚Ø¯ Ø§Ù„Ø¹Ù‚ÙŠØ¹ Ø¹Ø²ÙŠØ² Ø®Ù„Ø§Ù„
Ø§Ù„Ø§ÙŠØ§Ù… Ø§Ù„Ø«Ù„Ø§Ø«Ø© Ø§Ù^
Ø§Ù„Ø§Ø±Ø¨Ø¹Ø© Ø§Ù„Ø§Ø®ÙŠØ±Ø© Ù^Ø´Ø¯Ø¯Ø¹ Ø¹Ù„Ù‰ Ø§Ù†Ù‡ ÙŠÙ†Ø¨Ø²ÙŠ Ø¹Ù„Ù‰ Ø§Ù„Ø¹Ø±Ø§Ù‚
Ø§Ù„Ø§Ù„Ø¹Ø²Ø§Ù… Ø¨Ù‚Ø±Ø§Ø±Ø§Ø¹ Ù…Ø¬Ù„Ø³ Ø§Ù„Ø§Ù…Ù†
Ù„Ù„Ø¨Ø¯Ø¡ Ø¨Ù…Ø±Ø§Ø¬Ø¹Ø© Ø´Ø§Ù…Ù„Ø©" Ù„Ù„Ø¹Ù‚Ù^Ø¨Ø§Ø¹.
</P>
<P>
Ù^Ø§Ø¶Ø§Ù  Ø§Ù† "Ø¬Ù^Ø§Ø¨Ù‡ (Ø§Ù„ÙŠÙ^Ù…) ÙƒØ§Ù† Ø§Ù„Ø¬Ù^Ø§Ø¨ Ø§Ù„Ø°ÙŠ Ø³Ø¨Ù‚ Ù„Ù‡ Ø§Ù†
Ø§Ø¹Ø·Ø§Ù†ÙŠ Ø§ÙŠØ§Ù‡ Ù^Ù‡Ù^ Ø§Ù†Ù‡Ù…
Ù„ÙŠØ³Ù^Ø§ Ù…Ø³Ø¹Ø¹Ø¯ÙŠÙ† Ù„Ù„Ø¹Ù^Ø¯Ø© Ø¹Ù† Ù‚Ø±Ø§Ø±Ù‡Ù… Ø§Ù„Ø§ Ø§Ø°Ø§ Ù‚Ø§Ù… Ù…Ø¬Ù„Ø³
Ø§Ù„Ø§Ù…Ù† Ø¨Ø´Ù‰Ø¡ ÙŠØ¹Ø¹Ù„Ù‚ Ø¨Ø§Ù„Ø¹Ù‚Ù^Ø¨Ø§Ø¹".
</P>
<P>
Ù^Ù‚Ø¯ Ø§Ù^Ù‚Ù  Ø§Ù„Ø¹Ø±Ø§Ù‚ Ù ÙŠ 31 Ø¹Ø´Ø±ÙŠÙ† Ø§Ù„Ø§Ù^Ù„/Ø§ÙƒØ¹Ù^Ø¨Ø± Ø§Ù„Ù…Ø§Ø¶ÙŠ
Ø¹Ø¹Ø§Ù^Ù†Ù‡ Ù…Ø¹ Ø§Ù„Ù„Ø¬Ù†Ø© Ø§Ù„Ø®Ø§ØµØ©
Ù„Ù„Ø§Ù…Ù… Ø§Ù„Ù…Ø¹ØØ¯Ø© Ø§Ù„Ù…ÙƒÙ„Ù Ø© Ù†Ø²Ø¹ Ø§Ø³Ù„ØØ© Ø§Ù„Ø¯Ù…Ø§Ø± Ø§Ù„Ø´Ø§Ù…Ù„
Ø§Ù„Ø¹Ø±Ø§Ù‚ÙŠØ© (ÙŠÙ^Ù†ÙŠØ³ÙƒÙ^Ù…) Ø¨Ø¹Ø¯ Ø§Ù† ÙƒØ§Ù†
Ù‚Ø¯ ØØ¯ Ù…Ù† Ù†Ø´Ø§Ø·Ø§Ø¹Ù‡Ø§ Ù…Ù†Ø° Ø§Ù„Ø®Ø§Ù…Ø³ Ù…Ù† Ø§Ø¨/Ø§Ø²Ø³Ø·Ø³ Ø§Ù„Ù…Ø§Ø¶ÙŠ.
</P>
<P>
Ù^ÙŠØ´Ø¹Ø±Ø· Ø§Ù„Ø¹Ø±Ø§Ù‚ Ù„Ù„Ø¹Ø±Ø§Ø¬Ø¹ Ø¹Ù† Ù‚Ø±Ø§Ø±Ù‡ Ø§Ù† ÙŠØ¨ØØ« Ù…Ø¬Ù„Ø³ Ø§Ù„Ø§Ù…Ù†
Ø±Ù Ø¹ Ø§Ù„Ø¹Ù‚Ù^Ø¨Ø§Ø¹ Ø§Ù„Ù…Ù Ø±Ù^Ø¶Ø©
Ø¹Ù„ÙŠÙ‡ Ù…Ù†Ø° Ø¹Ø§Ù… 1990.
</P>
<P>
Ù ÙŠ Ø§Ù„Ù…Ù‚Ø§Ø¨Ù„ Ø§Ø´Ø¹Ø±Ø· Ù…Ø¬Ù„Ø³ Ø§Ù„Ø§Ù…Ù† Ø§Ù† ÙŠØ³Ø¹Ø§Ù†Ù  Ø§Ù„Ø¹Ø±Ø§Ù‚ Ø¹Ø¹Ø§Ù^Ù†Ù‡
Ù…Ø¹ Ø§Ù„Ù„Ø¬Ù†Ø© Ø§Ù„Ø®Ø§ØµØ© Ù‚Ø¨Ù„
Ø§Ø¬Ø±Ø§Ø¡ Ø§ÙŠ Ù…Ø±Ø§Ø¬Ø¹Ø© Ù„Ù„Ø¹Ù‚Ù^Ø¨Ø§Ø¹.
</P>
<P>
Ù^Ø§Ø´Ø§Ø± Ø¨Ø±Ø§ÙƒØ§Ø´ Ø´Ø§Ù‡ Ø§Ù„Ù‰ Ø§Ù†Ù‡ Ø§Ø¨Ù„Ø º Ø¹Ø²ÙŠØ² Ø¨Ù‚Ø±Ø§Ø± Ø§Ù„Ø§Ù…Ù…
Ø§Ù„Ù…Ø¹ØØ¯Ø© Ø³ØØ¨ Ø¬Ù…ÙŠØ¹ Ù…Ù^Ø¸Ù ÙŠÙ‡Ø§ Ø²ÙŠØ±
Ø§Ù„Ø¶Ø±Ù^Ø±ÙŠÙŠÙ† "Ø¹ØØ³Ø¨Ø§ Ù„Ø¶Ø±Ø¨Ø§Ø¹ Ø¹Ø³ÙƒØ±ÙŠØ© Ù…ØØ¹Ù…Ù„Ø© Ø¶Ø¯ Ø§Ù„Ø¹Ø±Ø§Ù‚".
</P>
</TEXT>
</DOC>
```

Figure 3: Fichier source codé en UTF-8 avec ses balises

Dans le corpus chaque phrase du document est délimitée par <P> et </P>, et le titre est délimité par les balises <HEADLINE> et </HEADLINE>.

4.3 Exécution de Lakhas

Notre système donne à l'utilisateur la possibilité d'introduire le taux de compression qu'il veut avoir pour le résumé, ce taux est fixé par défaut à 50%.

Nous avons produits plus de 300 résumés sur le corpus et nous avons obtenu d'excellents résultats qui seront présentés plus loin.

La Figure 4 affiche le fichier de sortie de Lakhas qui représente les parties qui composent un résumé du fichier source représenté dans la Figure 2, où l'identificateur du document est mentionné ainsi que les phrases avec leur position et score en ordre décroissant. Dans cet exemple, on voit que sur les 8 phrases du document, que la 1ère, 7ème, 2ème et 5ème phrase ont été sélectionnées pour former un résumé compressé à 50%, de plus on remarque que le titre ainsi que les expressions indicatives ont joué un rôle important dans la sélection de ces phrases.

Figure 4: fichier de sortie de Lakhas visualisé par un browser

Sur les différents résultats que nous avons obtenus, nous avons remarqué que les paragraphes sélectionnés comme résumés comportent une certaine cohérence dans le contenu.

4.4 Effets de variation des coefficients sur la fonction score globale

Dans le but de déterminer les paramètres qui influencent le plus les résultats, nous avons fait des expérimentations en faisant varier les valeurs des coefficients de la fonction score globale donnée par l'équation 1.

Cette expérience nous a montré que le titre avait énormément d'influence sur les résultats des résumés.

Sur une vingtaine de documents nous avons remarqué que dès que le titre n'est pas pris en considération, les phrases saillantes changent beaucoup, et souvent même, d'autres phrases sont extraites en premiers. Alors que dans les autres cas les deux premières phrases extraites sont souvent les mêmes.

Le Tableau 10 donne la moyenne des scores des phrases d'un texte suivant les différents paramètres pour 21 documents ainsi que la différence entre les deux premières phrases extraites.

coefficients				Poids phr1	Poids phr2	Poids phr3	Poids phr4	Poids phr5	diff poids (phr1-phr2)
position	titre	tfidf	cue						
1	1	1	1	18,167	12,811	10,231	8,178	6,774	5,357
1	1	1	0	17,677	12,729	10,126	8,106	6,774	4,948
1	1	0	1	17,190	12,000	9,474	7,500	6,111	5,190
1	1	0	0	16,714	11,905	9,368	7,429	6,111	4,810
1	0	1	1	3,559	2,129	1,897	1,804	1,749	1,431
1	0	1	0	3,030	1,910	1,794	1,744	1,688	1,119
1	0	0	1	2,524	1,333	1,053	1,071	1,000	1,190
1	0	0	0	2,000	1,000	1,000	1,000	1,000	1,000
0	1	1	1	16,453	11,646	9,152	7,175	5,774	4,807
0	1	1	0	16,011	11,491	9,075	7,104	5,774	4,519
0	1	0	1	15,476	10,857	8,368	6,500	5,111	4,619
0	1	0	0	15,048	10,667	8,316	6,429	5,111	4,381
0	0	1	1	1,683	1,054	0,849	0,794	0,749	0,629
0	0	1	0	1,104	0,868	0,777	0,735	0,688	0,236
0	0	0	1	0,667	0,190	0,053	0,071	0,000	0,476
0	0	0	0	0,000	0,000	0,000	0,000	0,000	0,000

Tableau 10: La moyenne des poids des 5 premières phrases suivant les paramètres (position, titre, tfidf, cue)

Comme nous pouvons le constater la moyenne des poids pour la 1ère phrase saillante varie entre 18,167 et 15,048 pour les cas où le titre est pris en considération et entre 3,559 et 0 dans le cas contraire, ce qui montre l'influence des titres dans la sélection des résumés.

4.5 Évaluation de Lakhas

L'évaluation de résumés produits par des systèmes de résumé de texte automatiques est un processus complexe. Les critères à prendre en considération doivent concerner les trois paramètres qui caractérisent un bon résumé : le taux de compression, le taux de rétention et la cohésion de l'extrait.

La tâche d'évaluation est normalement exécutée manuellement par des juges qui comparent subjectivement des résumés différents et choisissent le meilleur. Le problème de cette approche tient au fait que les juges qui exécutent la tâche d'évaluation ont souvent des idées très différentes sur ce qu'un bon résumé devrait contenir. Un autre problème avec l'évaluation manuelle est son coût en temps.

L'évaluation automatique se voit comme une alternative intéressante ce qui a ouvert de nouvelles perspectives pour l'évaluation de résumés.

ROUGE[1] (Recall-Oriented Understudy for Gisting Evaluation) représente un système automatique d'évaluation de résumés. Ce système a été développé par [Lin, 2004] et dont les scores calculés par cette méthode semblent bien corrélés avec l'évaluation humaine. Il inclut des mesures pour déterminer automatiquement la qualité d'un résumé en le comparant à d'autres résumés modèles créés par des humains. Les mesures comptent le nombre d'unités de recouvrement des n-grammes entre le résumé généré automatiquement et les résumés modèles, ces unités représentent les termes du texte pour lesquels on applique une lemmatisation par "Porter Stemmer", laquelle consiste à ôter une terminaison prédéfinie (la plus longue possible) au mot étudié, puis à ajouter une terminaison prédéfinie à la racine obtenue (exemple : predication → predicate, motoring → motor) [Porter, 1980].

Le NIST a décidé d'employer ROUGE dans la conférence d'évaluation de résumés de document (DUC 2004) à laquelle nous avons participé. Les résultats que nous avons obtenus sont présentés à la section 5.5.

Le traitement automatique de la langue arabe est encore jeune et souffre donc d'un manque de ressource, en particulier pour les évaluations de système. Nous n'avons pas d'échantillon de documents avec des résumés référence pour faire une appréciation plus crédible, néanmoins nous avons fait des comparaisons à deux systèmes de résumé de texte arabe.

Cette appréciation consiste à comparer notre système avec deux systèmes commerciaux Sakhr Arabic Summarizer de la société Sakhr Software (http://www.sakhr.com), considérée comme un leader dans le traitement automatique de la langue Arabe et Pertinence Summarizer de la société Pertinence.net fondée par A.Lehmam qui lit et condense des textes de 14 langues.

[1] site web de ROUGE : http://www.isi.edu/~cyl/ROUGE/

	Lakhas	Pertinence	Sakhr
Taux de compression	Paramétrable	Paramétrable	20 à 40 %
Mots clés	Oui	Oui	Non
Mots d'exclusions	Non	Oui	Non

Tableau 11:Caractéristiques des systèmes Lakhas, Pertinence et Sakhr

Notre expérimentation a porté sur un échantillon de 26 documents et dont le nombre de phrases varie entre 4 à 13 phrases par document avec une moyenne globale de 6 phrases.

Taux de compression		Lakhas/Pertinence	Lakhas/Sakhr	Pertinence/Sakhr
17%	1 phrase pertinente	73%	77%	65%
33%	2 phrases pertinentes	69%	58%	60%
50%	3 phrases pertinenets	73%	50%	47%

Tableau 12: Corrélations des systèmes Lakhas/Pertinence/Sakhr

Ce taux représente la couverture moyenne en phrases des résumés entre chaque système.

Par exemple pour une compression de 17% c'est à dire en résumant le texte en une phrase nous avons obtenu pour les 26 documents, les 19 mêmes résumés que ceux de Pertinence et 20 pour ceux de Sakhr. Et pour une compression de 33% à deux phrase, nous avons obtenu 10 résumés identiques à 100% que ceux de Pertinence et les 16 autres recouvraient 50% des autres résumés de Pertinence.

A noter qu'à partir des résumés de plus d'une phrase quand Lakhas et Pertinence n'avaient pas des résumés identiques à 100%, ils possédaient toujours des phrases communes dans les résumés pour tous les documents, ceci peut être dû au fait que les deux systèmes exploitent les titres comme mots-clés.

Cette première expérimentation modeste nous permet d'affirmer que Lakhas est assez compétitif avec les systèmes commerciaux. Il reste encore à augmenter la taille de l'échantillon et à utiliser d'autres types de documents.

ROUGE aurait été intéressant à utiliser pour comparer ces trois systèmes, mais pour cela, il aurait fallu disposer de résumés modèles en texte arabe.

4.6 Conclusion

Nous avons présenté l'approche utilisée pour la production automatique de résumés en arabe, en décrivant l'architecture de Lakhas, le corpus utilisé et les expériences que nous avons réalisées, les résultats que nous avons obtenus sont très bons. Toutefois pour la validation du système, nous devrions disposer de modèles pour faire des comparaisons. Une autre forme de validation sera la participation à des conférences d'évaluation comme TREC (Text REtrieval Conferences) ou DUC (Document Understanding Conferences) mais qui sont spécifiques aux résumés de textes arabes. Cette expérience sera décrite au prochain paragraphe.

5 Lakhas à DUC 04

Ce chapitre décrit l'adaptation de Lakhas pour notre participation à la tâche 3 dans DUC 2004 (Document Understanding Conferences), une Conférence d'évaluation annuelle dans le secteur de résumé de texte automatique organisé par l'Institut National Américain de Standards et de Technologie (NIST).

Cette tâche est nouvelle et consiste à générer des résumés très courts (~10 mots) de chaque document à partir d'entrées bruitées produites par traduction automatique (de l'arabe à l'anglais) de documents arabes provenant de l'Agence France Presse.

Le scénario proposé par les organisateurs de DUC était le suivant (Figure 5) : les textes arabes originaux sont d'abord traduits en anglais par un système de traduction automatique (TA). Le texte anglais résultant est utilisé comme source pour le système de résumé automatique afin d'obtenir un résumé très court (~75 bytes) du document en anglais. Deux systèmes de Traduction Automatique ont été employés: un de l'Institut des Sciences de l'Information (ISI) de l'université de la Californie du sud et l'autre développé par une équipe d'IBM.

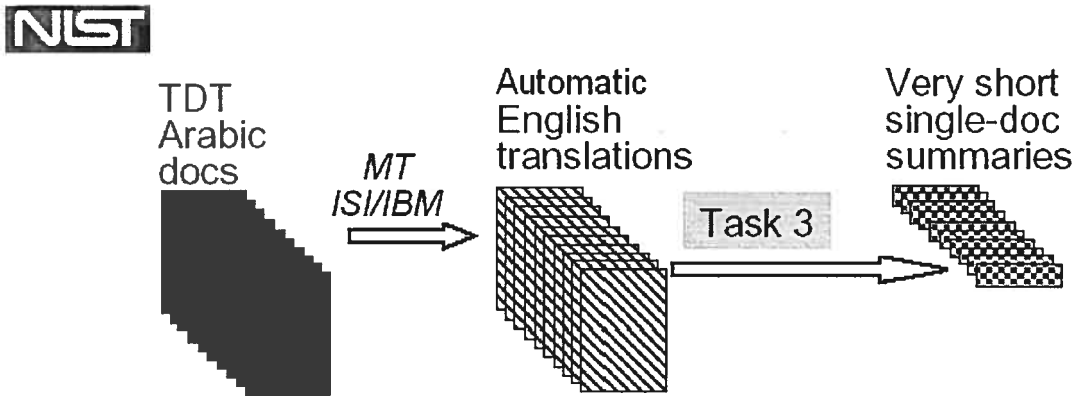

Figure 5: Scénario de NIST présenté par Paul Over [Over et Yen, 2004]

Après une étude préliminaire sur un petit échantillon d'exemples de documents traduits, nous avons constaté un certain nombre d'imperfections:

- Les textes anglais sont à peine compréhensibles sans les textes arabes correspondants.
- Les systèmes de Traduction Automatique ont souvent ignoré certaines informations importantes dans leur traduction. Par exemple, dans *The Agency said that Ibrahim, in the event at the level of cooperation and trade between Iraq and Saudi Arabia*, le verbe *appreciated* a été omis après le mot *event* ; même avec ce mot, le texte reste difficile à comprendre.
- Les systèmes de Traduction Automatique ont souvent traduit un même mot arabe par différents mots anglais par exemple منازل a été traduit par *home* dans une phrase et par *workers* dans une autre.

C'est pourquoi nous avons décidé de suivre un autre chemin pour notre participation à DUC2004: résumer le texte arabe directement et traduire seulement le texte résumé. Ainsi nous avons moins de texte à traduire mais plus important encore, nous travaillons directement sur les textes originaux, qui sont alors moins bruités dans le but d'améliorer les résultats, la Figure 6 schématise le scénario du RALI.

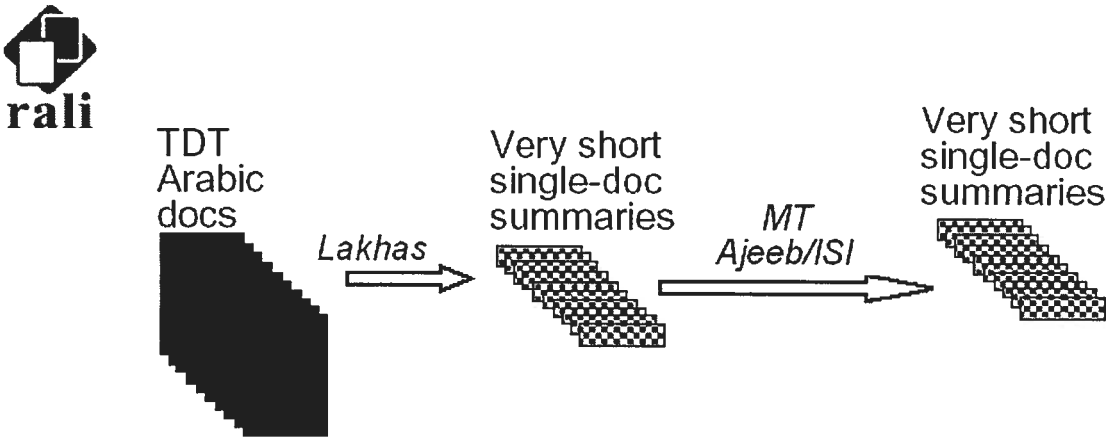

Figure 6: Scénario de RALI

Lakhas a été le premier système de résumé arabe à être formellement évalué et comparé avec des concurrents anglais dans une compétition internationale d'évaluation.

5.1 Structuration et normalisation des docset

Pour l'évaluation des résumés, NIST avait fourni aux participants 240 documents à résumer sous forme de 24 docsets, où chaque docset représentait un dossier de 10 documents concernant un même événement.

5.1.1 Structuration des données d'entrées

Etant donné que notre approche différait de celle proposée par le NIST dans DUC 2004, nous avons été contraint de construire et d'utiliser une table de correspondance entre docid fournis par LDC au DUC2004 et docid utilisé dans le catalogue Gigaword LDC2003T12 et de classer les docid selon les docset définis par NIST. Chaque docset représente un répertoire regroupant les documents appropriés.

5.1.2 Normalisation pour le traitement.

Comme les documents de Gigaword sont au format UTF-8, nous les avons convertis au format ISO-8859-6, afin qu'on puisse les traiter par notre système.

5.2 Méthode appliquée pour l'extraction des phrases pertinentes

Pour extraire la phrase la plus pertinente d'un document, nous avons utilisé une combinaison de quatre méthodes en considérant l'équation 1 donnée en section 3.1.7. dont les coefficients a_i ont été tous fixés à 1.

Cette étape permet de sélectionner la phrase la plus pertinente, elle est suffisante pour des résumés courts, la Figure 7 présente le nombre de mots dans les résumés produits par Lakhas par rapport aux textes sources.

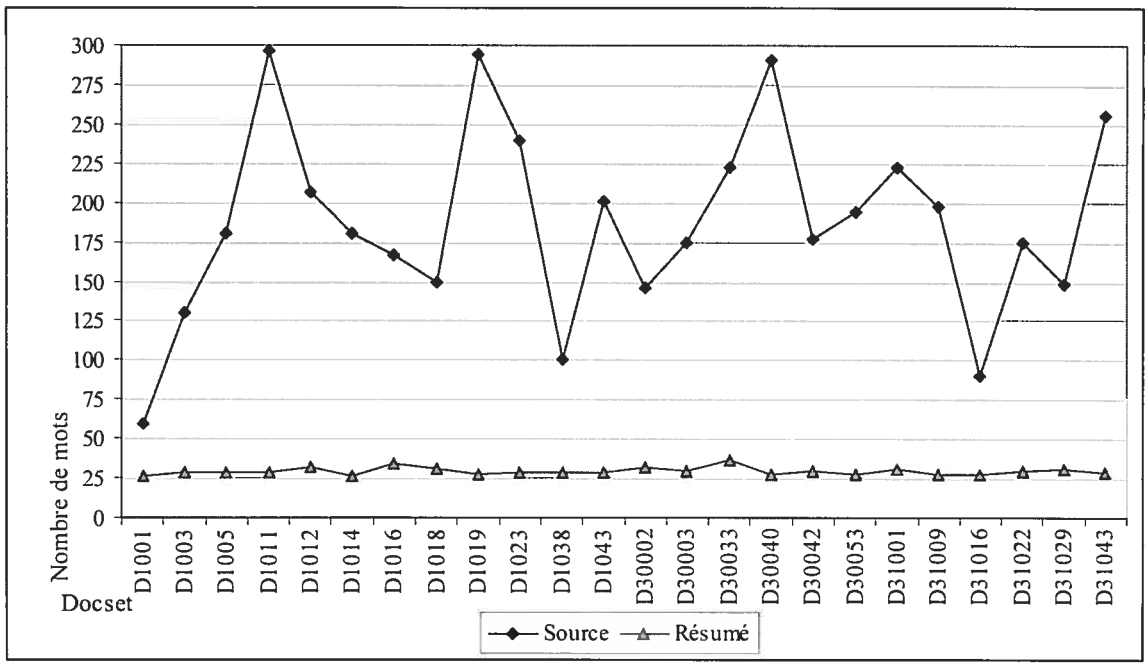

Figure 7: Moyenne de nombre de mots arabes par docset pour texte source et résumés

Comme nous pouvons le voir dans le graphique, les résumés produits par l'extraction ont environ 29 mots arabes en moyenne, mais l'évaluation à DUC a imposé que les résumés aient environ 10 mots anglais, nous supposons que la tâche consistait plutôt à générer des titres qui résument les articles. Pour satisfaire cette contrainte et réduire nos résumés, nous avons dû développer des procédures spécifiques de compression.

La Figure 8 schématise le processus suivi ainsi que les extensions apportées à Lakhas dans le but de participer à DUC.

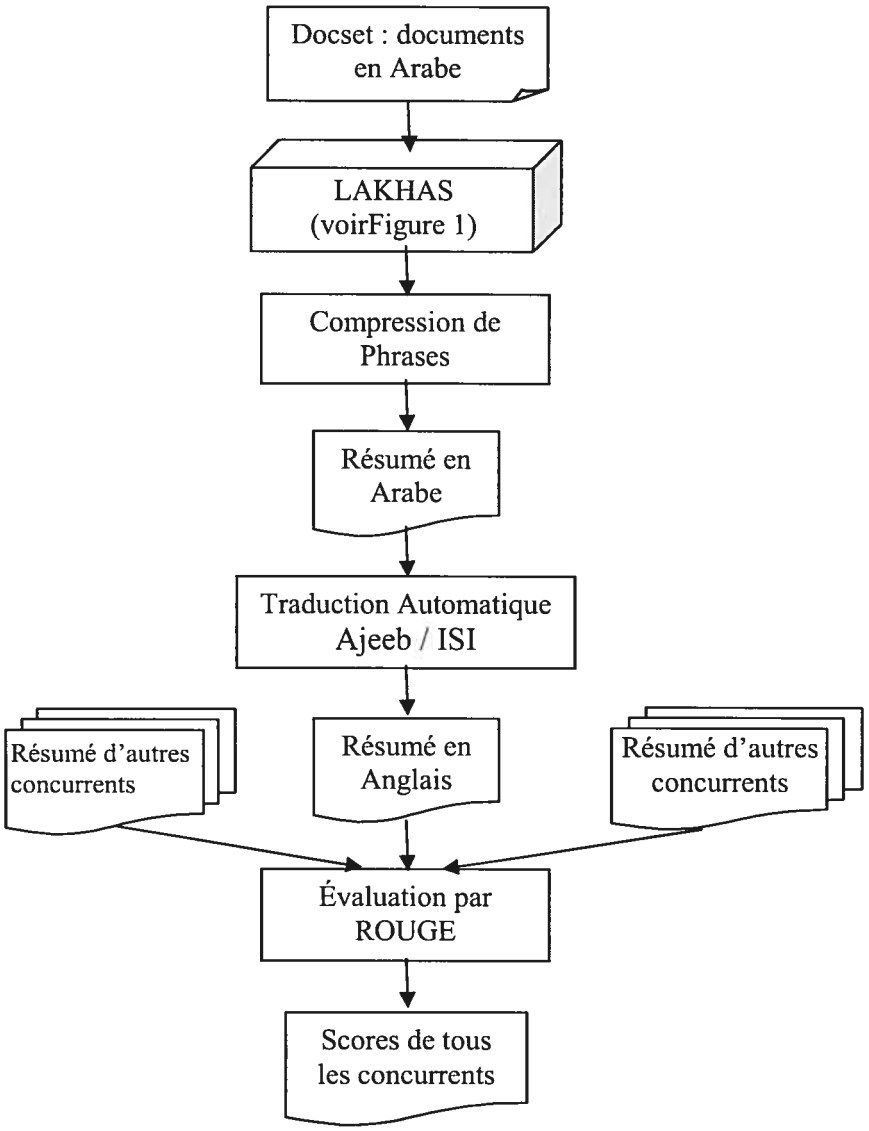

Figure 8: Extesions de Lakhas à DUC 2004

5.3 Méthodes appliquées pour la réduction de phrase.

Afin de réduire le nombre de mots dans les phrases sélectionnées comme résumé, nous avons appliqué quatre sortes de réductions.

5.3.1 Substitution de nom

Afin de réduire le nombre de mots qui représentent une personnalité politique ou sportif. Le principe est d'associer un nom représentatif (nom ou fonction) à chaque personnalité, le Tableau 13 donne un exemple de substitution.

Désignation de personnalité en Arabe	Désignation de personnalité en Français
رئيس دولة الامارات العربية المتحدة الشيخ زايد بن سلطان آل نهيان	**Le président** de l'Etat **des Emirats Arabes** unis Sheikh Zayed Bin Sultan Al-Nahyan
الامين العام للامم المتحدة كوفي انان	Secrétaire général des Nations Unis **Kofi Annan**
وزير الداخلية المصري حبيب العادلي	**Le ministre de l'intérieur égyptien** Habib el-Adly

Tableau 13: Exemple de substitution où nous ne conservons que les mots gras soulignés

5.3.2 Suppression de mots non expressifs

On supprime des mots qui n'ajoutent pas d'information substantielle tels que des jours de la semaine ou les mois, les chiffres écrits en lettres, les adverbes, quelques conjonctions de subordination, etc.. Le but est de réduire le nombre de mots sans pour autant influencer le sens de la phrase. On tolère donc quelques erreurs syntaxiques qui de toutes façon seront modifiées lors de la traduction.

Le Tableau 14 donne quelques exemples de mots non expressifs

Nature des mots à supprimer	*Mots en arabe*	*Mots en français*
Les jours et les mois	يوم السّبت, يوم الأحد , في يناير, فبراير	samedi, dimanche,…, janvier, février,…
Les fréquences et repères de temps	أمس, غدًا, في الصّباح, المساء حتى , اثناء	hier, demain, matin, soir,… jusqu'à,.. Pendant'..
Les adjectifs démonstratifs	هذا, هذه, هؤلاء,	ce, cette, ces,…
Les pronoms démonstratifs et relatifs	هذا, هذه, ذلك, الذي, ذلك, الذي,	ceci, cela, celle, qui, que, lequel, lesquels,…
Les conjonctions de subordination	مثلا , مثل, متى	comme, lorsque, …
Les mots d'accentuation (d'intensité)	أساسي , حتمًا, رسمي	principalement, nécessairement, officiel, …
Les mots d'indication (signalisation)	أكد, أعلن, أبلغ, قال , يعتبر	confirmé, annoncé, déclaré, dit, considère
Les mots de répétition	لايزال	encore,…
Les dérivées du verbe être	كانت يكون	était, qui sera, …
Les nombres en lettres	اول , تاني …	premier, deuxième,…
Adjectifs indéfinis	بعض	quelques,
Les mots de contraste	لكنه	Mais,
Les mots objectifs (but, déduction)	حيث ,	Pour, afin

Tableau 14: exemple de catégorie de mots arabe à supprimer et leur traduction en français

5.3.3 Suppression de parties de phrases à partir de frontières

Dans cette étape il ne s'agit pas de segmenter un texte mais plutôt de réduire une phrase en essayant de garder son sens. On tronque la phrase à partir de certaines frontières.

Les frontières ont été déterminées à partir des exemples traités, et concernent la ponctuation, les conjonctions de coordination ou de subordination et dans certains cas des mots connecteurs:

- de coordination و , مع , كما, مثل *et, avec, comme, comme a,...*
- d'indication خلال, منذ, *depuis, pendant,...*
- d'explication التي, حول *qui, celle, a propos,...*
- de causalité لهذا *dû,...*

Le connecteur و *et* peut jouer différents rôles et pose une certaine difficulté pour l'identifier ou connaître sa fonction dans la phrase.

- Il peut désigner un lien ou une relation entre deux entités, exemple

  ...الازمة بين الامم المتحدة و العراق ***la crise entre l'ONU et l'Irak*** ,

  ...يبرز وحدة المسلمين والاقباط في المجتمع المصرى ***montre la cohésion des musulmans et les copts dans la société égyptienne.***

- Il peut représenter une conjonction entre deux propositions mais on ne peut reconnaître s'il fait partie du mot ou s'il désigne une conjonction car il est collé au mot

Le Tableau 15 donne quelques exemples d'interprétation de mots débutant par و où il est difficile de reconnaître si le و fait partie du mot ou pas.

	mot du texte	mot décomposé	traduction	Mot compacté	Traduction
1	وجود	"و" "جود"	et générosité	"وجود"	Existence
2	وهم	"و" "هم"	et angoisse	"وهم"	Illusion
3	وقائع	"و" "قائع"	et ∅	"وقائع"	Evénements
4	وجهت	"و" "جهت"	et ∅	"وجهت"	dirigée
5	ورفض	"و" "رفض"	et a refusé	"ورفض"	∅

Tableau 15:exemples d'interprétation possible de mots débutant par و.

Un dictionnaire pourrait aider à éviter la troncature de la phrase dans les cas 3 et 4 du Tableau 15. Le fait de tronquer à partir du و *et* va donner une information incomplète ce qui rendrait la phrase dans certains cas difficile à comprendre.

5.3.4 Suppression des constructions de discours indirect

Étant donnée la nature du corpus, nous avons dégagé certains types motifs comme indiqués dans le Tableau 16:

	Motif en arabe	Motif en français
1	**R** ان ...X أعلن	X a déclaré .... que **R**.
2	**R** ان ...X ذكر	X a mentionné ... que **R**
3	**R** ان ...X أفاد	X a reporté ... que **R**
4	**R** بان ...X أوضح	X a clarifié ... que **R**
5	**R** انه ...X أعلن	X a déclaré .... que cela est **R**.

Tableau 16: Quelques modèles de motifs avec leur traduction en français

Nous avons décidé de détacher les termes neutres extérieurs au thème principal et nous avons choisi de ne garder que le reste de la phrase (défini par **R**) qui pourra rester significatif même en supprimant les sources d'information, le Tableau 17 donne un exemple de suppression de constructeur de discours indirect.

Texte source	*Traduction en français*
أعلن نائب رئيس الجمهورية العراقي طه ياسين رمضان اليوم الأحد أن العراق يرفض التراجع عن قراره بوقف التعاون مع مفتشي نزع الاسلحة الدوليين قبل تلبية مطالبه.	*Le vice-président irakien Taha Yassin Ramadan **a déclaré** aujourd'hui dimanche **que** l'Irak refuse le retrait de sa décision sur l'arrêt de coopération avec les inspecteurs internationaux de désarmement avant d'effectuer ses requêtes*

Tableau 17: Exemple de suppression de construction de discours indirect en appliquant le modèle 1 du tableau 16. Le texte souligné représente le segment à retenir dans le résumé.

Un des inconvénients de l'utilisation de constructeurs de discours indirect pour la compression de phrases est la génération d'information incomplète, il arrive que dans certains textes au lieu de mentionner une entité nommée, on fait référence à l'orateur, par exemple dans le cas de *pays* comme entité, au lieu de dire *le vice président Irakien a déclaré aujourd'hui dimanche que l'Irak refuse la coopération avec les inspecteurs*, nous avons plutôt *le vice président Irakien a déclaré*

aujourd'hui dimanche que son pays refuse la coopération avec les inspecteurs, l'utilisation du premier patron du Tableau 16 dans ce cas nous donnerait *son pays refuse la coopération avec les inspecteurs*, qui est une phrase incomplète où une information importante (quel pays?) est omise. Un processus de résolution anaphorique pourrait remédier à cette insuffisance.

Le mot انه utilisé comme frontière dans le 5ème modèle du Tableau 16 peut avoir deux interprétations:

i. Comme pronom démonstratif neutre *ceci/cela*, mais n'influe pas sur les résultats.

ii. Comme conjonction de subordination *que* avec le pronom personnel *il* où ه désigne le *il*, ce qui rendrait la phrase incomplète après la réduction, présence d'anaphore sans antécédent, le Tableau 18 donne un exemple où une information importante est omise dans le résumé (qui est candidat ?).

Texte source	*Traduction en français*
اعلن رئيس الحكومة الجزائرية مولود حمروش فرانس برس انه مرشح للانتخابات الرئاسية لوكالة	Le Premier ministre algérien Mouloud Hamrouche ***a déclaré*** à l'agence France Presse ***qu'il*** est candidat aux élections présidentielles.

Tableau 18: Exemple d'information incomplète en utilisant le 5ème modèle du Tableau 16, le texte souligné représente le segment retenu dans le résumé.

En appliquant les quatre méthodes de réduction précédentes, nous avons été capable de réduire les résumés de moitié, comme le montre la colonne 5 du Tableau 19. Ces quatre méthodes de réduction nous ont permis d'avoir des résumés d'environ 15 mots en arabe, tout en gardant les informations les plus importantes.

DOCSET	Source	Résumé	Résumé très court	Résumé/Source	Résumé très court/Source	Traduction de Ajeeb	Traduction de ISI
D1001	59	26	14	44%	23%	20	16
D1003	130	29	12	22%	9%	18	15
D1005	180	29	15	16%	8%	24	22
D1011	297	29	17	10%	6%	28	24
D1012	207	32	13	15%	6%	18	14
D1014	180	26	11	15%	6%	17	13
D1016	167	34	18	20%	11%	27	24
D1018	149	31	13	21%	9%	19	16
D1019	294	27	15	9%	5%	24	19
D1023	239	29	15	12%	6%	22	18
D1038	100	29	14	29%	13%	20	19
D1043	201	28	13	14%	6%	20	16
D30002	146	32	16	22%	11%	21	18
D30003	174	30	17	17%	10%	28	22
D30033	223	36	17	16%	8%	25	20
D30040	291	27	15	9%	5%	20	16
D30042	177	30	14	17%	8%	18	17
D30053	194	27	15	14%	8%	22	18
D31001	222	31	13	14%	6%	21	16
D31009	197	27	16	14%	8%	24	20
D31016	90	27	16	30%	18%	25	19
D31022	175	30	15	17%	9%	20	20
D31029	148	31	14	21%	10%	22	17
D31043	256	29	12	11%	5%	18	17
Moyenne	**187**	**29**	**15**	**16%**	**8%**	**21**	**18**

Tableau 19: Moyenne du nombre de mots arabes par docset pour texte source et résumés, les deux dernières colonnes donnent le nombre de mots dans les textes anglais correspondants.

5.4 Processus d'extraction appliqué à DUC 2004

L'entrée de Lakhas était le nom du docset qui représentait le nom du répertoire regroupant les documents en arabe à parcourir et la sortie consistait en un seul fichier qui regroupait les résumés du docset.

Pour la sortie, nous étions obligé d'utiliser les mêmes références que celles de NIST. A cet effet, nous avons utilisé un fichier de correspondance "nist-gigaword.txt" avec

la fonction "giga2nist", qui a permis de produire un dossier regroupant les résumés des documents de chaque docset, séparés par des balises désignant les mêmes références que ceux du NIST, la Figure 9 donne un exemple de fichier de sortie de Lakhas pour DUC 2004.

```
PERDOC DOCREF="AFA19981216.1800.0259" DOCSET="d100it"
عمليات القصف ضد العراق تتم خصوصا بواسطة صواريخ عابرة تطلق من حاملة الطائرات الاميركية وقاذفات بي-52
/PERDOC

PERDOC DOCREF="AFA19981217.1800.0336" DOCSET="d100it"
عمليات القصف الكثيفة على العراق قد تتوقف ابتداء من
/PERDOC

PERDOC DOCREF="AFA19981218.1400.0147" DOCSET="d100it"
طالب وزير الخارجية الالماني في بون بوقف سريع لعمليات القصف على العراق.
/PERDOC

PERDOC DOCREF="AFA19981218.1400.0140" DOCSET="d100it"
عدة مستشفيات عراقية ولا سيما مستشفى صدام اكبر مستشفى في العراق اصيبت في القصف الاميركي البريطاني للعاصمة العراقية
/PERDOC

PERDOC DOCREF="AFA19981219.1800.0253" DOCSET="d100it"
مسؤول وزارة الدفاع الاميركية (البنتاغون) سيوصون بيل كلينتون بوقف القصف على العراق.
/PERDOC

PERDOC DOCREF="AFA19981219.1000.0076" DOCSET="d100it"
مدنيا عراقيا قتلوا نتيجة القصف الاميركي والبريطاني على العراق تم تشييع جثثهم في بغداد 68.
/PERDOC

PERDOC DOCREF="AFA19981219.1000.0083" DOCSET="d100it"
لقى 68 مدنيا عراقيا على الاقل مصرعهم منذ من جراء القصف الاميركي والبريطاني على العراق
/PERDOC

PERDOC DOCREF="AFA19981220.1800.0152" DOCSET="d100it"
الاف الاشخاص قتلوا او جرحوا خلال الايام الاربعة من القصف الجوي للعراق.
/PERDOC
```

Figure 9: Exemple de Sortie de Lakhas pour DUC

Afin de comparer nos résultats à ceux d'autres équipes à DUC, nous avons traduit nos résumées arabes par Ajeeb (http://english.ajeeb.com) un système de traduction arabe-anglais commercialisé sur le web.

Pour traduire notre fichier avec la TA d'Ajeeb, nous avons dû :

- Encoder les documents dans le format de Windows CP-1256.
- Produire une page Web et envoyez son URL à Ajeeb.
- Transformer la page Web traduite dans un fichier XML conformément à la DTD de DUC.

5.5 Résultat à DUC 2004

La Figure 10 liste les différents participants à la tâche 3 de la compétition Document Understanding Conferences pour 2004.

Le déroulement de cette compétition suivait un agenda défini préalablement par le NIST et consistait en :

- le 16 Février, les données tests étaient disponibles sur le site Web du NIST.
- le 1 Mars, tous les participants soumettent leurs résumés à NIST pour l'évaluation.
- le 26 Mars, les résultats de l'évaluation sont rendus disponibles sur le site du NIST[1].

Task 3: Participants and runs

Sysid	Priority	Run	Group
UMD.BBN.Trimmer	1	7	U.Md/BBN
UMD.BBN.Trimmer	2	8	
CL	1	12	CL Research
CL	2	13	
LARIS.2004	1	20	Laris Labs
LARIS.2004	2	21	
MEDLAB_Fudan	1	37	Fudan U.
MEDLAB_Fudan	2	38	
MEDLAB_Fudan	3	39	
CLaCDUCTape2	1	58	Concordia U.
CLaCDUCTape2	2	59	
Lakhas0001	1	74	U. Montreal
webcl2004	1	82	ISI/USC
lcc.duc04	1	105	LCC
lcc.duc04	2	106	
uofo	1	112	U. Ottawa
uofo	2	113	
ie_ucd_iirg	1	133	U. College Dublin
ie_ucd_iirg	2	134	
UofM-MEAD	1	141	U. Michigan
UofM-MEAD	2	142	
UofM-MEAD	3	143	
webcl2004	2	151	ISI/USC

Priority 1 (required): input = IBM/ISI automatic translations
2 (required): input = Manual translations
3 (optional): input = automatic translations + relevant Eng. documents

Figure 10: Liste des participants pour la tâche 3, présentée par P. Over à DUC 2004 [Over et Yen, 2004]

[1] http://duc.nist.gov/duc2004/active/results/ROUGE/t3.rouge.out.tab

Lors de cette compétition, des intervenants ont même testé leur système sur des traductions manuelles (Priority 2, Figure 10). Ce fut le cas pour les systèmes 8, 13, 21, 38, 59, 106, 113, 134 et 142.

Le Tableau 20 donne un extrait des données sources ainsi que les différents résumés du document AFA19981218.0000.0001 du docset d1001t. On remarque que le résumé du concurrent 142 est un extrait de la traduction manuelle, de plus les constructions de discours indirect *announced that* sont gardées dans le résumé.

Traduction manuelle document source	Washington 12-18 (AFP) - The American State Department announced that Russia recalled her ambassador to the United States "for consultation" due to the bombing operations on Iraq.
Traduction Automatique par ISI du document source	Washington 12-18 (AFP) - The US State Department announced that Russia had led to the ambassador in the United states "consultation" because of the bombings on Iraq.
Résumé manuel	Russia recalled its US ambassador because of bombings in Iraq
Résumé concurrent 141	The US State Department announced that Russia had led to the ambassador in|
Résumé concurrent 142	The American State Department announced that Russia recalled her ambassador|
Résumé concurrent 3 (baseline)	Washington 12-18 (AFP) - The US State Department announced that Russia had|
Résumé Lakhas (traduit par Ajeeb)	Russia called its ambassador in the United States for the consultation beca|

Tableau 20: Données sources et résultats concernant le documents AFA19981218.0000.0001 du docset d1001t, pour les systèmes 141,142 et Lakhas.

NIST a évalué les résumés anglais avec ROUGE (section 4.5) en utilisant 4 modèles de résumés comme références construits par des humains à partir des traductions manuelles.

Le Tableau 21 compare les résultats de Lakhas (LKS) avec ceux des autres participants à DUC2004

ID	ROUGE-1		ROUGE-2		ROUGE-3		ROUGE-4		ROUGE-L		R-W-1.2	
Titre	*0.351*		*0.154*		*0.074*		*0.036*		*0.316*		*0.188*	
model	*0.395*		*0.147*		*0.064*		*0.027*		*0.344*		*0.200*	
142	0.218	7	0.076	1	0.029	1	0.010	1	0.201	5	0.126	4
134	0.259	1	0.047	9	0.011	12	0.002	15	0.220	1	0.129	2
LKS	**0.236**	**5**	**0.052**	**6**	**0.016**	**8**	**0.003**	**8**	**0.207**	**2**	**0.125**	**6**
8	0.255	3	0.075	2	0.026	2	0.009	2	0.207	3	0.127	3
59	0.255	2	0.071	3	0.023	3	0.006	3	0.206	4	0.126	5
...												
3	0.137	24	0.029	20	0.009	17	0.002	14	0.116	24	0.074	24

Tableau 21: Score de Rouge pour quelques systèmes ainsi que leurs rangs

La dernière ligne du Tableau 21 (participant 3) présente les performances du système minimal (baseline) sur les traductions automatiques des textes. Ces scores sont très mauvais du fait qu'on tronque à 75 bytes les résumés ce qui ne laisse que les 10 premiers mots pour l'évaluation, or les débuts des premières phrases sont généralement des annonces sans information pertinente (voir Tableau 20).

Étant donné la nature de la tâche (résumé en 10 mots), les participants n'étaient pas autorisés à utiliser les titres comme résumés. La première ligne du Tableau 21 donne une idée sur les scores obtenus si on considère comme résumé les traductions manuelles des titres.

Comme nous pouvons le constater sur la Figure 11 Lakhas (participant 74) a de bons résultats (classé au $5^{ème}$ ou $6^{ème}$ rang) comparé à d'autres systèmes malgré que nous ayons suivi un chemin totalement différent avec un autre système de traduction automatique que celui qui avait été utilisé pour traduire les documents arabes originaux.

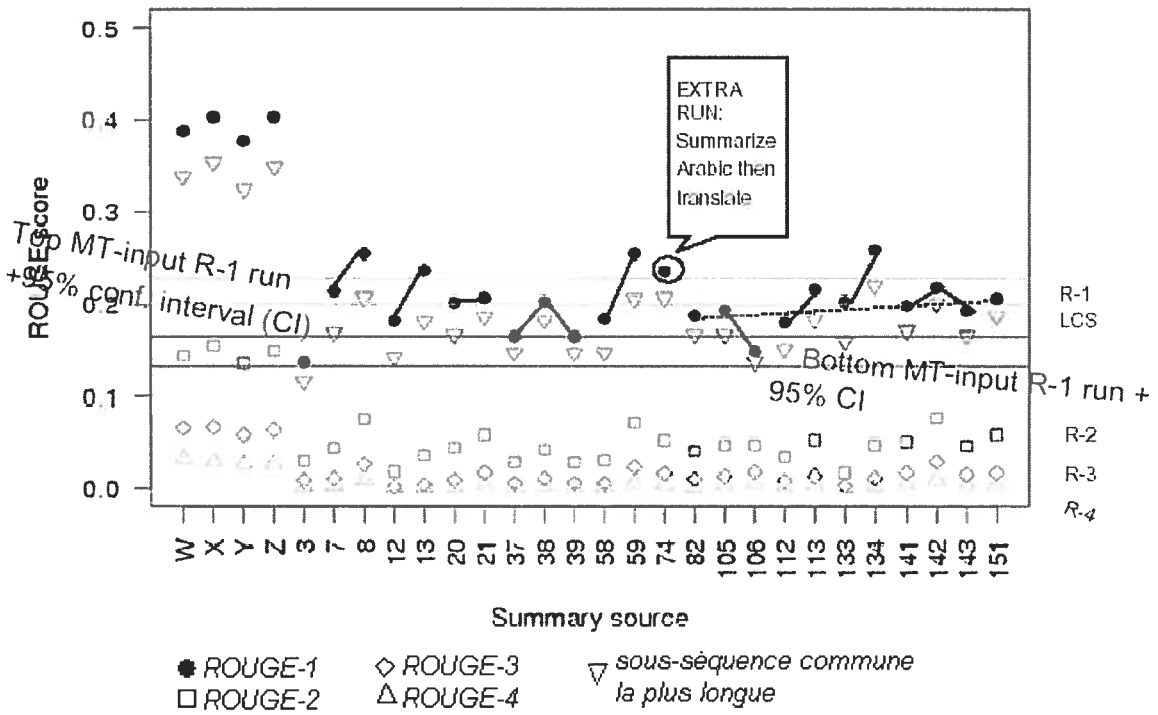

Figure 11: Score de Rouge par participant, présenté par Paul Over à DUC 2004 [Over et Yen, 2004], les abscisses représentent les participants et les modèles W' X, Y, Z, et les différent scores de ROUGE en ordonnées.

Après la compétition grâce à la collaboration de Franz Och à ISI, nous avons obtenu une traduction anglaise de nos résumés arabes avec le système ISI, un des deux qui avaient été utilisés pour traduire les document originaux.

Nos scores sont alors devenus les meilleurs de tous (voir Tableau 22) où la ligne de LKS-ISI représente les nouvelles valeurs.

ID	ROUGE-1		ROUGE-2		ROUGE-3		ROUGE-4		ROUGE-L		R-W-1.2	
model	*0.395*		*0.147*		*0.064*		*0.027*		*0.344*		*0.200*	
LKS-ISI	**0.297**	**1**	**0.084**	**1**	**0.029**	**1**	**0.009**	**3**	**0.256**	**1**	**0.153**	**1**
142	0.218	7	0.076	2	0.029	2	0.010	1	0.201	6	0.126	4
134	0.259	2	0.047	10	0.011	13	0.002	16	0.220	2	0.129	2
LKS	***0.236***	***6***	***0.052***	***7***	***0.016***	***9***	***0.003***	***9***	***0.207***	***3***	***0.125***	***6***
8	0.255	4	0.075	3	0.026	3	0.009	2	0.207	4	0.127	3
59	0.255	3	0.071	4	0.023	4	0.006	4	0.206	5	0.126	5
...												
3	0.137	25	0.029	21	0.009	18	0.002	15	0.116	25	0.074	25

Tableau 22: Nouveau score de Rouge en introduisant la traduction de ISI

De même que sur la Figure 11, on peut voir le nouveau score ROUGE-1 obtenu sur les résumés traduis par ISI est imposant, même dans le cas où les autres concurrents ont utilisé les traductions manuelles.

Les procédures de réduction que nous avons ajoutées à Lakhas, en particulier la suppression des constructions de discours indirect, ont permis à Lakhas de se distinguer des autres participants par la préservation de plus d'information pertinente.

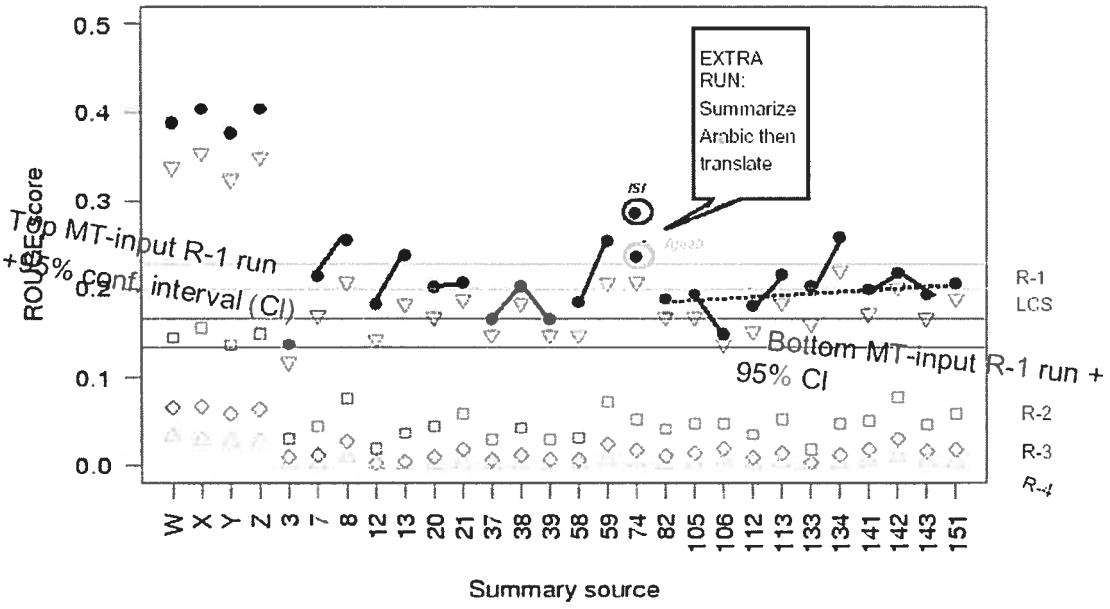

Figure 12: Introduction du nouveau score ROUGE

5.6 Impact de la traduction de l'arabe à l'anglais

A première vue, les traductions faites par Ajeeb paraissaient appropriées. Mais en observant les résultats donnés par ROUGE de LKS par rapport à LKS-ISI, nous avons été un peu surpris et nous en avons déduit que les erreurs de traduction étaient principalement responsables de certains de nos mauvais scores.

Pour expliquer ce phénomène de différence des scores, nous analyserons dans ce qui suit les traductions de Ajeeb et ISI en tenant compte du fait que ROUGE utilise des modèles de résumé et établit une troncature à partir du 10ème mot.

5.6.1 Erreur de traduction

Dans les traductions faites sur *Ajeeb* :

- Cas des mots avec erreur de saisie : il fait la translittération du mot (suppose que c'est un nom propre, Tableau 23),

Mot saisi	Traduction	Mot correct	Traduction	L'erreur
لشلروع	Lshlroa	للشروع	Pour entreprendre	Permutation du " ل "
الفلسطينين	Alflstinin	الفلسطينيين	Les Palestiniens	Manque " ي "

Tableau 23: Exemple de traduction de Ajeeb de mots inconnus.

- Cas de mots ambigus: comme تطلق est traduit incorrectement par Ajeeb en *divorce,* alors qu'il peut avoir deux interprétations *lancer* ou *divorcer*. Pour ce mot, les diacritiques auraient pu donner la bonne interprétation sans même prendre en considération le contexte. تُطَلِّقْ= *elle divorce*, تُطْلَقْ = *elle est lancée*
- L'article du pluriel indéfini dans l'arabe ي qui est collé aux mots pour désigner l'article indéfini *des*, entraîne une ambiguïté du fait que ي peut aussi désigner l'adjectif possessif *mon* comme le montre le Tableau 24.

Texte en arabe	Traduction de Ajeeb	Traduction correcte
مسؤلي وزارة الدفاع الاميريكية(البنتاغون)	My official the US Department of Defense (the Pentagon).	Officials at the American Defense Department (The Pentagon).

Tableau 24: Exemple de traduction de Ajeeb pour l'article indéfini ي

- L'absence de majuscule dans la langue arabe influence aussi sur la traduction en particulier pour les noms propres, comme pour le cas du mot بون, qui désigne la ville Bonn en Allemagne, a été traduit par *différence.*

5.6.2 Développement des mots lors de la traduction

La langue arabe, par sa nature compositionnelle des mots, accroît souvent le nombre de mots lors de la traduction, le Tableau 25 donne un exemple de ce développement lors de la traduction par deux systèmes de traduction automatiques, un vers l'anglais et l'autre vers le français :

1 Mot en arabe	MT (Ajeeb) 4 mots en anglais	MT (Reverso) 5 mots en français
فذكرناهم	Then we mentioned them	Alors nous les avons mentionnés

Tableau 25: Exemple de développement d'un mot arabe lors de sa traduction vers l'anglais puis vers le français

Les principaux facteurs qui influencent ce développement sont :

- l'article ال. *le, les, la* …, exemple العمليات se décompose en 2 mots *les opérations*
- Les adjectifs possessifs ه , ه , هم , … *son, ses, leur,…*, exemple عملهم se décomposent en 2 mots *Leur travail*
- Les adverbes بـ , لل *par, en,.sur la,…*, exemple في القصف للعاصمة *dans le bombardement sur la capitale*
- Les conjonctions و *et, exemple* والتقى *et il a rencontré*
- Les pronoms personnels : en arabe lors de la conjugaison, le sujet est inclus dans le verbe. Il n'est donc pas nécessaire de précéder le verbe conjugué de son pronom. Ce qui va développer le mot en sujet et verbe lors de la traduction سيوصون ils conseilleront.

Le Tableau 19 donne le nombre de mots pour les résumés originaux en arabe ainsi que le nombre de mots de la traduction correspondante. Le fait que des résumés ont été tronqués à 75 bytes avant leur évaluation a également influencé sur la baisse de nos scores. Nos résumés arabes avaient 15 mots qui ont été augmentés à 21 mots anglais lors de la traduction par Ajeeb et 18 lors de la traduction par ISI, dont beaucoup ont été mal tronqués avant évaluation (voir le Tableau 26).

De plus nous avons remarqué que pour des mots inconnus ou incorrectement orthographiés, ISI les ignorait tandis que Ajeeb les gardait avec un meilleur effort de translittération.

5.6.3 Comparaison des traductions avec un model de référence

Un autre paramètre qui a influencé les résultats est le fait que les mots traduits par ISI sont généralement identiques aux mots utilisés dans les modèles de référence tandis que ceux de Ajeeb sont souvent synonymes (voir Tableau 26). On remarque aussi que dans ISI les bi-grammes sont beaucoup plus présents (*to_participate*, *will_support*, ...) que dans Ajeeb.

Model	Traduction de Ajeeb / Traduction de ISI	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	R-W-1.2
King Hussein nearly **finished** *chemotherapy* **treatments** at American Mayo Clinic	Al-Malik Hussain ended the fourth stage from a *chemotherapy* from origin six|stages	0.13	0.03	0.00	0.00	0.13	0.09
	King Hussein finished the fourth phase of the chemical **treatment** of the six|stages	0.42	0.15	0.03	0.00	0.39	0.22
Nelson Mandela arrives to participate in **annual** Gulf States **summit**	Nelson Mandela arrived to the United Arab Emirates for the participation in| the annual summit to the Gulf countries	0.35	0.11	0.06	0.04	0.33	0.20
	Nelson Mandela arrived in the Emirates to participate in the **annual summit**| of the Gulf	0.60	0.28	0.16	0.00	0.55	0.32
Cohen confident Gulf *countries* will support **"appropriate action" against Iraq**	The Gulf Arab *countries* will offer the support to the doing of a suitable w|ork against Iraq	0.37	0.06	0.00	0.00	0.34	0.21
	Gulf Arab states will support for **"appropriate action" against Iraq,**	0.55	0.27	0.13	0.04	0.53	0.32

Tableau 26: Les scores de ROUGE pour des traductions de phrases par Ajeeb et ISI. Les mots soulignés n'ont pas été pris en compte pendant l'évaluation en raison de la troncature. Italique/Gras est pour des mots trouvés dans la traduction d'Ajeeb/ISI et également dans le modèle.

ROUGE semble être un outil très intéressant pour l'évaluation de résumés, mais il dépend des modèles de référence utilisés.

5.7 Conclusion

Dans ce travail nous avons participé à DUC 2004 mais en suivant une autre approche que celle proposée par le NIST. A cet effet, nous avons utilisé quelques techniques d'extractions, qui sont en fait une combinaison de quatre méthodes appliquées à des documents en langue arabe. Nous avons ensuite ajouté quatre procédés de réduction.

Cette expérimentation est intéressante par son approche et ses résultats, elle a montré qu'il est plus pratique de traiter des données qui respectent en général une certaine structure ; de plus, le travail sur les textes en langue originale donne accès à plus d'informations fiables.

Dans les traitements automatiques de la langue où l'ambiguïté est omniprésente que ça soit en traduction, en recherche d'information ou en résumé automatique, faisant appel à des techniques basées sur la statistique, la linguistique et les règles, la moindre omission d'informations clés influence négativement les résultats.

Les résultats obtenus sont très bons mais leurs évaluations par ROUGE semblent influençables par les systèmes de traductions, ceci est dû au fait que ROUGE utilise comme référence des modèles générés par des humains qui ont travaillé sur les textes traduits manuellement.

6 Conclusion et perspectives

Nous avons décrit Lakhas, dont l'objectif est d'élaborer un système capable de résumer automatiquement des textes en langue arabe. Nous avons étudié cette langue d'un point de vue informatique et fait ressortir les traits communs et les particularités par rapport aux langues latines plus souvent traitées en TAL.

Les méthodes d'extraction se sont avérées adaptables et nous avons pu faire nos expérimentations sur le corpus Gigaword qui regroupe un ensemble de nouvelles journalistiques en arabe.

Nous avons aussi vérifié l'apport de certains paramètres sur les résumés extraits par notre système où les titres (mots clés) semblaient jouer un rôle important.

Malheureusement, il n'existe pas jusqu'à ce jour de méthode de validation, sauf le fait de participer à des compétitions d'évaluation ou de disposer de ressource nécessaire comme données de jugements pour calculer les précision et rappel. Néanmoins nous avons pu comparer notre système avec deux systèmes commerciaux dont les résultats nous ont paru concurrentiels. De plus nous avons participé à une compétition d'évaluation de résumé en anglais (DUC 2004) où nous avons obtenu de bons et même les meilleurs résultats dépendamment des systèmes de traduction. Ceci confirme le choix de résumer les textes dans leur langue originale plutôt qu'avec leur traduction.

On pense que Lakhas peut encore être plus performant au niveau de taux de compression en lui intégrant le module de *suppression des constructions de discours indirect* mais en ajoutant un traitement qui permet de corriger l'existence d'anaphore sans antécédent.

Un axe qui pourrait être intéressant est l'exploitation des caractéristiques de la langue arabe (règles de dérivation + agglutination) en vue de la génération de texte pour les résumés automatiques.

7 Bibliographie

[Aljlayl et Frieder, 2002] M. Aljlayl and O. Frieder, On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach, *In 11th International Conference on Information and Knowledge Management (CIKM), November 2002, Virginia (USA), pp.340-347.*

[Amini et Gallinari, 2002] M. R. Amini, P. Gallinari : Apprentissage numérique pour le résumé de texte, *Les Journées d'Étude de l'ATALA, Le résumé de texte automatique : solutions et perspectives, décembre 2002, Paris (France) (http://www.atala.org/je/021214/Amini.pdf)*

[Attia, 2000] M. Attia, A large-scale computational processor of the Arabic morphology, *A Master's Thesis, Cairo University, (Egypt) 2000.*

[Baloul et al., 2002] S. Baloul, M. Alissali, M. Baudry, P. Boula de Mareüil: Interface syntaxe-prosodie dans un système de synthèse de la parole à partir du texte en arabe, *24es Journées d'Étude sur la Parole, 24-27 juin 2002 Nancy, pp.329-332.*

[Berri, 1996], J. BERRI : Mise en œuvre de la méthode d'exploration contextuelle pour le résumé automatique de textes. Implémentation du système SERAPHIN, *Actes du colloque de CLIM'96, Montréal, pp.128-135.*

[Chali et Pinchak, 2001] M.B. Yllias Chali, C. J. Pinchak : Text Summarization Using Lexical Chains, *Proceedings of the Document Understanding Conference (DUC 2001), New Orleans, USA pp135-140.*

[Chalabi, 2001], A. Chalabi : Sakhr Web-based Arabic<>English MT engine, *ACL/EACL 2001 Workshop on Arabic Language Processing, Toulouse July 2001.(http://www.elsnet.org/arabic2001/chalabi.pdf)*

[Chen et Gey, 2002] A. Chen and F. Gey : Building an Arabic Stemmer for Information Retrieval. *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002).* National Institute of Standards and Technology, Nov 18-22, 2002, *pp631-640.*

[Darwish, 2002] K. Darwish: Building a Shallow Arabic Morphological Analyzer in One Day. *Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, USA. pp. 47-54.*

[Darwish, 2003] K. Darwish : Probabilistic Methods for Searching OCR-Degraded Arabic Text, *Doctoral dissertation, University of Maryland, 2003*

[Débili et al., 2002] Débili F., Achour H., Souici E. : La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique, *Correspondances de l'IRMC, N° 71, juillet-août 2002, pp. 10-28.*

[Desclés et al., 2001] J.P. Desclés, J.L. Minel, E. Cartier, G. Crispino, S. Ben Hazez, A. Jackiewicz : Résumé automatique par filtrage sémantique d'informations dans des textes, *Présentation de la plate-forme FilText, Technique et Science Informatiques, 2001, n°3, pp. 369-395.*

[Edmundson, 1969] H. P. Edmundson: New methods in automatic abstracting, *Journal of the Association for Computing Machinery (ACM), vol. 16 N°2 April 1969, pp. 264-285.*

[Elhadad et Barzilay, 1997] M. Elhadad & R. Barzilay : Using Lexical Chains for Text Summarization, *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid 1997, pp. 10-17.*

[Farzindar, 2003] A. Farzindar, Résumé automatique de textes juridiques, *Proposition de projet doctoral, Université de Montréal, Septembre 2003.*

[Hammou et al., 2002] Hammo B., Abu-Salem H., Lytinen S., Evens M., QARAB: A Question Answering System to Support the Arabic Language, *Workshop on Computational Approaches to Semitic Languages. ACL 2002, July 2002, Philadelphia, PA, pp. 55-65.*

[Hernandez et Grau, 2002] N. Hernandez et B. Grau, Analyse Thématique du Discours: segmentation, structuration, description et représentation, *Colloque International sur le Document Électronique (CIDE), 20-23 octobre 2002, Hammamet, Tunisie, pp. 277-288.*

[Ishikawa et al., 2001] Ishikawa, K., Ando, S., Okumura, A.: Hybrid Text Summarization Method based on the TF Method and the Lead Method. *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization. Tokyo. Japan. March 2001. pp.5-219-5-224.*

[Jaruskulchai et Kruengkrai, 2003] C. Jaruskulchai and C. Kruengkrai: A Practical Text Summarizer by Paragraph Extraction for Thai. *The Sixth International Workshop on Information Retrieval with Asian Languages (IRAL-2003), July 7, 2003, Sapporo Japan, pp. 9-16.*

[Kadri et Benyamina, 1992] Y. Kadri, A.Benyamina, Système d'analyse syntaxico-semantique du langage arabe, *mémoire d'ingénieur, université d'Oran Es-sénia, 1992*

[Kiraz, 1996], G. A. Kiraz : Analysis of the Arabic Broken Plural and Diminutive, In Proceedings of the 5th International Conference and Exhibition on Multi-Lingual Computing (*ICEMCO96), Cambridge, UK*

[Larkey et al., 2002] Larkey L. S., Ballesteros L. and Connell M., Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis, *In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, August 2002, pp. 275-282.*

[Leclerc, 2000] J. Leclerc, L'aménagement linguistique dans le monde, *http://www.tlfq.ulaval.ca/axl/monde/famarabe.htm*

[Lehmam, 2000] A. Lehmam Résumé de texte automatique: des solutions opérationnelles, *La Tribune des Industries de la Langue, de l'Information Électronique et du Multimédia, OFIL, Paris, pp.50-58.*

[Lin, 2004] C-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, *Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain, pp. 74-81.*

[Luhn, 1958] P. H. Luhn: The Automatic Creation of Literature Abstracts, IBM Journal April 1958, *pp. 159-165.*

[Minel et al., 2001) Minel J.-L., Ferret O., Grau B., Porhiel S., Repérage de structures thématiques dans des textes, *Actes de TALN 2001, 02-05 juillet 2001, Tours, France, pp. 163-172.*

[Nobata et Sekine, 2003] C. Nobata & S. Sekine, Results of CRL/NYU System at DUC-2003 and an Experiment on Division of Document, *Proceedings of the Document Understanding Conference (DUC 2003), Edmonton, Canada. pp.79-84*

[Ouersighni, 2001] Ouersighni R. 2001. A major offshoot of the DIINAR-MBC project: AraParse, a morphosyntactic analyzer for unvowelled Arabic texts, *ACL/EACL 2001 Workshop on Arabic Language Processing, Toulouse July 2001, pp. 9-16.*

[Over et Yen, 2004] P. Over, J. Yen, An Introduction to DUC 2004 Intrinsic Evaluation of Generic New Text Summarization Systems, *Proceedings of the Document Understanding Conference (DUC 2004), Boston (USA), May 6-7 2004, pp. 1-21.*

[Pardo et al., 2002] T.A.S. Pardo, L.H.M. Rino, M.G.V. Nunes, Extractive summarization: how to identify the gist of a text, *International*

Information Technology Symposium - I2TS 2002, Florianópolis-SC, Brazil, 01-05 October 2002, pp.245-260.

[Perrin et Lehmam,] D. Perrin et A. Lehmam: Le traitement de la documentation face à la production démultipliée d'information, *La Tribune Des Industries de la Langue et du Multimédia, périodique n33-34*

[Porter, 1980] M. F. Porter, An Algorithm for Suffix Stripping. *Program, 14: 1980, pp.130-137.*

[Saggion, 2000] H. Saggion, Génération automatique de résumés par analyse sélective, *Thèse de Ph.D en Informatique, Université de Montréal, août 2000.*

[Saggion et Lapalme, 2000] H. Saggion, G. Lapalme, Selective Analysis for the Automatic Abstracting: Evaluating Indicativeness and Acceptability, *Proceedings of Content-Based Multimedia Information Access, RIAO'00, Paris (France) 2000, pp. 747-764.*

[Strzalkowski et al., 1998] T. Strzalkowski, J. Wang and B. Wise, Summarization-based Query Expansion in Information Retrieval, *Proceedings of 36th Annual Meeting of the ACL, Montreal 1998, V. 2, pp. 1258-1264.*

[Torres et al., 2001] Torres-Moreno J.M, Velázquez-Morales P. et Meunier J.G., Cortex : un algorithme pour la condensation automatique des textes. *Colloque Interdisciplinaire en Sciences Cognitives ARCo'2001, Décembre 2001, Lyon (France).*

[Xu et al., 2002] Xu J., Fraser A., Weischedel Ralph, Empirical Studies in Strategies for Arabic Retrieval, *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002), August 11-15, 2002, pp. 269-274.*