REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Hadj Lakhdar – Batna
Faculté des Sciences
Département d'Informatique

MÉMOIRE
Présenté par
Tahar DILEKH

Pour obtenir le grade de
Magister
Spécialité : **S**ystème d'**I**nformation et de **C**onnaissance (**SIC**)

Implémentation d'un outil d'indexation et de recherche des textes en arabe

---

Soutenue publiquement le 28 /09 /2011 devant le jury formé de :

Pr. Mohammed BENMOHAMMED	Professeur	**Président**	Université de Constantine
Dr. Abdelmadjid ZIDANI	M.C.	**Rapporteur**	Université de Batna
Dr. Ali BEHLOUL	M.C.	**Co-Rapporteur**	Université de Batna
Pr. Azzeddine BILAMI	Professeur	**Examinateur**	Université de Batna
Dr. Brahim BELATTAR	M.C.	**Examinateur**	Université de Batna

ABSTRACT

Arabic, one of the six official languages of the United Nations, is the mother tongue of more than 300 million people[1]. Becoming a center of research and commercial development, the importance of the domain of information retrieval (IR) is due to the essential need for such tools to people in the Net era. The number of Internet users in 2002 was about 4.4 million which represents about 1.5 % of the population of the Arab world[2]. Few search engines are available for the growing number of Arabic Internet users; however many effort are being deployed,

Arabic is a highly inflected language and has a complex morphological structure. The information retrieval (IR) on Arabic texts requires the basic form of the word (root or lemma); therefore stemming process is necessary. This process can be defined as a process of removing all affixes (prefixes, infixes and/or suffixes) from words in order to bring them to their roots or lemmas.

Morphological complexity of Arabic language makes particularly developing software for natural languages process difficult. In Semitic languages like Arabic, the majority of names, adjectives and verbs are derived from few thousand roots by adding new letters, for example, the words مكتبة (library), كتاب (book) كتب (pounds) كتب (he wrote), and نكتب (writing), the root كتب [Wig98].

Every natural language has its own characteristics and features. Thus, it is difficult to follow the same pattern of stemming and apply the same techniques for all languages. A technique of stemming could be relevant to a language, but could not be effectively applied to other languages. There are several techniques used for word stemming; these include technical dictionaries, morphological analysis, deletion of affixes, statistics, and translation.

In this work, we proposed a hybrid method that incorporates three different techniques so as the Arabic stemming process resolve problems in connection with each mentioned technique.

These three techniques are: removal of affix given by Kadri [Kad08], dictionaries, and morphological analysis

These techniques require some adjustments to be relevant for use. Each technique is fitted and adapted individually to solve practical problems related to itself.

Therefore, the main contribution of this experiment was to demonstrate the effectiveness of the *hybrid method* compared to other methods, and the choice of removing the suffix before prefix during the operation of Arabic stemming process. For example

[1] http://www.frcu.eun.eg/www/homepage/cdc/cdc.htm. Accédé le 12/12/2008.
[2] http://sonomabusiness.com/archives/2002-09-column-levini.html. Accédé le 05/12/2008.

Word	Root	Suffix – Préfixe	Prefix - Suffix
ألامهم	ألم	ألام	مها
Their pain	pain	Pains	error

Key Words: Information retrieval, lemmatization, Arabic language.

RESUME

L'arabe, une des six langues officielles des Nations Unies, est la langue maternelle de plus de 300 millions de personnes[1]. Le domaine recherche d'information (RI) arabe, devenu un centre de la recherche et du développement commercial est du à la nécessité essentielle de tels outils pour des personnes dans l'ère électronique. Le nombre d'internautes arabophones en 2002 était environ 4.4 millions, environ 1.5% de la population du monde arabe[2]. Mais, de l'autre côté de la réalité, peu de moteurs de recherche sont mis à la disposition des utilisateurs arabophones, bien que les efforts soient en marche pour servir le nombre croissant d'utilisateurs.

L'Arabe est une langue fortement flexionnelle qui a une structure morphologique complexe. La recherche d'information sur le texte arabe exige la forme de base du mot (racine ou lemme) pour être la plus pertinente, donc le processus de lemmatisation est nécessaire. La lemmatisation peut être définie comme un processus qui consiste à retirer tous les affixes (préfixes, infixes, ou/et suffixes) des mots pour ramener ces derniers à leurs lemmes ou racines.

La complexité morphologique de la langue arabe rend particulièrement difficile le développement des applications pour le traitement en langue naturelle. Dans les langues sémitiques comme l'arabe, la plupart des lemmes de nom, d'adjectif, et de verbe sont dérivés de quelques mille racines par l'insertion de nouvelles lettres, par exemple, les mots مكتبة (bibliothèque), كتاب (livre), كُتُبٌ (livres), كَتَبَ (il a écrit), et نَكْتُبُ (nous écrivons), de la racine كتب [Wig98].

Chaque langue naturelle a ses propres caractéristiques et dispositifs. Ainsi, il est difficile de suivre la même configuration de lemmatisation et d'appliquer les mêmes techniques pour toutes les langues. Une technique de lemmatisation pourrait être pertinente à une langue, mais ne peut pas être effectivement appliquée à d'autres langues. Il existe plusieurs techniques utilisées pour la lemmatisation des mots. Celles-ci incluent, des techniques de dictionnaires, d'analyse morphologique, de suppression des affixes, de statistiques, et de traduction.

Dans ce travail, nous avons proposé une méthode hybride qui incorpore trois techniques différentes pour que la lemmatisation arabe résolve les problèmes liés à chaque technique précédente.

Ces trois techniques sont: suppression d'affixe proposée par Kadri [Kad08], dictionnaires, et analyse morphologique.

[1] http://www.frcu.eun.eg/www/homepage/cdc/cdc.htm. Accédé le 12/12/2008.
[2] http://sonomabusiness.com/archives/2002-09-column-levini.html. Accédé le 05/12/2008.

Ces techniques ont besoin d'une certaine adaptation pour être pertinentes pour l'utilisation. Chaque technique est adaptée individuellement pour résoudre les problèmes pratiques liés à elle-même.

La contribution principale de ce travail concerne la démonstration de l'efficacité de la *méthode hybride* comparée aux autres méthodes, et le choix de l'enlèvement des suffixes avant les préfixes pendant l'opération de lemmatisation Arabe. Par exemple

Mot	Racine	Suffixe – Préfixe	Préfixe - Suffixe
ألامهم	ألم	ألام	مها
Leurs douleurs	douleur	douleurs	Erreur

Mots clés : Recherche d'information, lemmatisation, langue arabe.

REMERCIMENTS

Je tiens tout d'abord à exprimer ma profonde gratitude au Docteur Ali BAHLOUL, maître de conférence au département d'informatique de l'Université de Batna, pour m'avoir encadré avec une grande compétence, pour sa disponibilité, son soutien, ses conseils qui m'étaient et me sont très utiles, ainsi que ses encouragements qui m'ont permis de mener à bien ce travail.

J'exprime ma profonde reconnaissance au Docteur Abdelmadjid ZIDANI, maître de conférence au département d'informatique de l'Université de Batna, pour son aide, son soutien, ses conseils, et pour m'avoir encadré.

Je remercie monsieur le professeur Mohammed BENMOHAMMED de m'avoir fait l'honneur d'examiner ma thèse et de présider le jury de ma soutenance. Je présente également mes remerciements aux professeurs Brahim BELATTAR et Azzeddine BILAMI qui ont accepté d'examiner ce travail. Je leur suis pleinement reconnaissant pour leur participation à ce jury.

Je remercie toutes les personnes qui ont participé de manière directe ou indirecte à la concrétisation de ce travail et plus particulièrement mes amis Tayeb MERZOUG, Kamel AMROUSSI qui m'ont aidé au cours de la correction de ce mémoire. Qu'ils trouvent ici l'expression de ma reconnaissance.

Je profite également de cette opportunité pour remercier mes très chers parents qui m'ont toujours soutenu, mes sœurs et mes frères pour leur soutien sans faille, leurs encouragements et sans lesquels rien n'aurait été possible.

Je remercie plus particulièrement ma femme pour sa présence et son soutien indéfectible même dans les moments difficiles au cours de mon parcours. Elle a su m'accompagner dans cette expérience scientifique qu'est l'élaboration de ce mémoire de magistère, sans oublier les fruits de ma vie mes deux beaux petits enfants, ZEINEB et REDHA.

Je veux, bien évidemment, remercier les personnes avec lesquelles j'ai travaillé durant ce parcours car sans le travail collectif, rien n'était possible : Guezouli, Noui, Belloula, Sassi, Bennzzar, Naili, Maache, Rabehi et bien d'autres.

TABLE DES MATIERES

LISTE DES TABLEAUX

LISTE DES FIGURES

INTRODUCTION GENERALE

Comme nous le savons, tout le monde a besoin d'informations dans sa vie quotidienne. Nous avons besoin de l'information pour prendre les meilleures décisions possibles. Dans chacune de nos activités personnelles, les décisions sont requises et l'information est nécessaire pour soutenir ces décisions. L'information est nécessaire dans presque tous les domaines de la pensée et l'action humaine.

Récemment, l'information numérique et les nouvelles formes de la technologie d'information sont devenues le centre d'intérêt de notre société. La croissance de la technologie d'information (TI) a contribué à la disponibilité de l'information enregistrée. Cela conduit à la nécessité de concevoir de nouvelles méthodes pour accéder à ce volume croissant des informations produites. Par conséquent, si elle doit être utilisée d'une manière plus économique et fructueuse, alors elle doit être organisée. La recherche d'information est le processus par lequel les informations (ou les documents qui les contiennent) sont stockées et mises à la disposition des utilisateurs, et la récupération de celles qui sont pertinentes aux besoins des utilisateurs [Gim01].

Les systèmes de recherche d'information SRI [Sal83] sont conçus pour faciliter l'accès aux informations stockées. En outre, ces systèmes sont concernés par la représentation, le stockage et l'organisation d'informations. Le système RI se compose d'un ensemble d'informations, d'un ensemble de requêtes et d'un mécanisme permettant de déterminer quelles sont les informations les plus pertinentes qui répondent aux besoins d'une requête.

Les techniques du traitement automatique des langues permettent d'extraire des textes des informations plus riches que de simples unités lexicales. Ces informations de nature morphologique, syntaxique et sémantique ont été partiellement utilisées en RI pour améliorer les méthodes d'appariement, les représentations des contenus des documents et requêtes et le processus de recherche.

Les langues naturelles se fondent sur des règles grammaticales, syntaxiques et morphologiques. Le niveau de difficulté et de complexité dépend de la langue elle-même. L'arabe est une langue hautement flexionnelle et a un besoin particulièrement aigu d'une normalisation et lemmatisation efficaces [Kho99] [Lar01] [Dar02] [Che02].

La langue arabe pose plusieurs défis au traitement du langage naturel (NLP) en raison de sa morphologie riche. Dans cette langue, le traitement morphologique devient particulièrement important pour la Recherche d'Information (RI), parce que la RI doit déterminer une forme appropriée de mots comme index. La lemmatisation des mots arabes a été un sujet central de nombreuses recherches en RI. Khoja [Kho99] a tenté de trouver les racines des mots arabes qui sont beaucoup plus abstraits que les lemmes. Il a supprimé d'abord les préfixes et suffixes, puis a tenté de trouver la racine de la forme dépouillée. McNamee [McN02] a utilisé la correspondance

n-grammes de plusieurs longueurs pour indexer les mots qui génèrent un index de grande taille. Les lemmatiseurs légers développés par Larkey [Lar01], Darwish [Dar02] et Chen [Che02] choisissent quelques préfixes et suffixes pour tronquer les mots. Cette dernière approche est inspirée du processus de lemmatisation de l'Anglais. Puisqu'elle donne le meilleur rendement, elle est largement utilisée aujourd'hui dans la RI.

Malgré ces travaux, il est encore peu clair quel type de lemmatisation est approprié pour la recherche d'information arabe. Une recherche plus poussée des effets de la lemmatisation sur l'efficacité de la RI est donc nécessaire. Ceci constitue un objectif important parmi plusieurs objectifs spécifiques de ce travail.

1. Objectifs

1.1. Objectif général

L'objectif général de ce travail est de développer un outil d'indexation et de recherche de la langue Arabe (un prototype).

1.2. Objectifs spécifiques

Les objectifs spécifiques de ce travail sont les suivants:

- Étudier les différents algorithmes de lemmatisation qui ont été développés pour la langue Arabe;
- Étudier les propriétés de la langue Arabe afin de se familiariser avec les différents aspects de la langue (par exemple, comment segmenter le texte Arabe).
- Sélectionnez le corpus (texte Arabe) de test et le préparer pour le traitement ;
- Collecte d'une liste d'affixes (préfixe et suffixe) utilisés dans l'Arabe;
- Collecte d'une liste de mots vides de la langue Arabe;
- Collecte d'une liste de modèles (AOUZANE) de la langue Arabe ;
- Collecte d'une liste de racines (DJOUDOUR) de la langue Arabe;
- Proposer un nouvel algorithme pour la lemmatisation d'Arabe ;
- Évaluer le nouvel algorithme afin de mesurer son efficacité;
- Conclure en spécifiant les perspectives pour des études plus approfondies.

2. Méthodologie

Comme l'étude de la morphologie de la langue constitue un composant important dans la recherche, une vue générale de littérature est faite pour rassembler des informations et comprendre le sujet [Kho99]. En outre ; des enseignants de la langue arabe sont consultés sur la morphologie d'Arabe.

Le corpus de texte est l'une des ressources nécessaires dans la RI. Un texte de bonne taille peut montrer le comportement morphologique raisonnable d'une langue. Alors, la sélection de texte est un composant important dans le développement d'un lemmatiseur. Afin d'atteindre le but de cette recherche, nous collectons et concevons un corpus « Al-Khat Alakhdar »[1]. Ce dernier a été recueilli pour étudier la langue Arabe. Par conséquent, le même corpus a été utilisé comme un échantillon représentatif pour étudier le comportement morphologique des mots et pour développer un lemmatiseur d'Arabe.

Après une étude détaillée de la morphologie de la langue, nous proposons des prétraitements relatifs à la normalisation morphologique de certains caractères, ces traitements touchent principalement la lemmatisation des mots arabes. Nous collectons des mots vides (Stop Word), des affixes (sachant que pour choisir les affixes, nous nous sommes basés sur la nouvelle méthode de Kadri [Kad08]) et nous les examinons en premier lieu. Les caractéristiques des affixes ont été ensuite utilisées pour orienter le développement du lemmatiseur. Enfin, nous développons un lemmatiseur sur la base de notre corpus de test.

Le test de notre lemmatiseur est fait sur une partie de notre corpus. Ensuite, nous avons évalué ce lemmatiseur, en utilisant une technique de comparaison entre cinq méthodes de lemmatisation :

- Enlever le préfixe ensuite le suffixe sans faire passer le mot sur les modèles;
- Enlever le suffixe ensuite le préfixe sans faire passer le mot sur les modèles;
- Enlever le préfixe ensuite le suffixe et faire passer le mot sur les modèles;
- Enlever le suffixe ensuite le préfixe et faire passer le mot sur les modèles;

La dernière méthode est basée sur l'hybridation entre les méthodes précédentes pour choisir le meilleur lemme.

Les erreurs sont manuellement comptées. Cette technique nous permet d'évaluer la performance de notre lemmatiseur.

La contribution principale de ce travail était donc la démonstration de l'efficacité de la méthode hybride comparée aux autres méthodes, et le choix de l'enlèvement des suffixes avant les préfixes pendant l'opération de lemmatisation de l'arabe.

Comme indiqué par Alemayehu et Willet [Ale02], le lemmatiseur est également applicable à d'autres applications de traitement de langue naturelle, qui nécessitent une analyse morphologique telles que les études de fréquence des mots, les correcteurs d'orthographe, les dispositifs d'analyse de texte, etc. Par conséquent, toute recherche et/ou de développement qui tienne compte de la morphologie Arabe pourraient bénéficier de ce travail.

---

[1] http:// www.greenline.com.kw. Accédé le 14/09/2008.

3. Organisation du mémoire

Ce mémoire est structuré en cinq chapitres

- Dans le premier chapitre, nous présentons les mécanismes de la RI, les concepts clés et les modèles de base sur lesquels repose la recherche d'information. Ainsi que la problématique de la recherche d'information arabe.
- Le second chapitre explique la morphologie de la langue arabe avec sa complexité et tente de déterminer les défis de cette langue dans la recherche d'information.
- Dans le troisième chapitre, nous exposons les différents prétraitements appliqués au corpus avant l'indexation.
- Des discussions sur la lemmatisation et le développement de l'algorithme de lemmatisation pour le texte arabe sont exposées dans le quatrième chapitre. La compilation de la liste des mots vides et la liste des affixes sont présentées dans ce chapitre. L'approche utilisée pour développer le lemmatiseur et les motivations de son choix sont également des parties de la discussion dans ce chapitre.
- Le cinquième chapitre explique les différentes étapes d'implémentation et d'expérimentation d'un outil d'indexation et de recherche ainsi que les résultats obtenus.

Finalement nous concluons notre travail avec l'ouverture des perspectives essentielles qui sont susceptibles de l'enrichi d'avantage et affiner nos contributions.

CHAPITRE 1 : LA RECHERCHE D'INFORMATION

1. Introduction:

Le but de la recherche d'information (RI) est de développer des systèmes permettant de retrouver les documents pertinents à une requête d'utilisateur, à partir d'une base de documents volumineuse. La notion de pertinence est très complexe. De façon générale, dans un document pertinent, l'utilisateur doit pouvoir trouver les informations dont il a besoin. C'est sur cette notion de pertinence que le système doit juger si un document doit être retourné à l'utilisateur comme réponse. Pour cela, il est important d'effectuer certains prétraitements sur les documents et les requêtes, c'est ce qu'on appelle l'indexation. Cette opération vise à construire une structure d'index qui permet de retrouver très rapidement les documents incluant des mots demandés.

L'indexation consiste donc à associer à chaque document (ou à chaque requête) un descripteur (également nommé index) formé de l'ensemble des termes d'indexation extraits de son contenu.

Pour déterminer si la représentation d'un document correspond à celle de la requête, on doit développer un processus d'évaluation. Différentes méthodes d'évaluation ont été développées, en relation avec la représentation de documents et de requêtes. C'est cet ensemble de représentation et de méthode d'évaluation qu'on appelle un modèle de RI; ils permettent :

- d'offrir une interprétation aux descripteurs en donnant une représentation interne des textes et des questions basée sur les termes d'indexation ;
- de définir les stratégies à adopter pour comparer les représentations des documents et des requêtes. Leur comparaison donne lieu à un score qui traduit leur degré de ressemblance ;
- de proposer éventuellement des méthodes de classement des résultats retournés à l'utilisateur.

Une fois les représentations des documents et des requêtes mises en correspondance, le système retourne à l'utilisateur la liste des documents répondant à sa requête. Une partie de ce chapitre y est consacrée.

Ne pas oublier aussi que la croissance rapide du World Wide Web (WWW) accompagnée d'une explosion des outils Web multilingues, de Web spider, d'indexation, et de recherche influent effectivement sur le développement dans le domaine de RI arabe.

2. Processus de recherche d'information

Une information est une donnée dont l'utilisateur a besoin pour résoudre un problème particulier. Il exprime donc son besoin sous forme de requête.

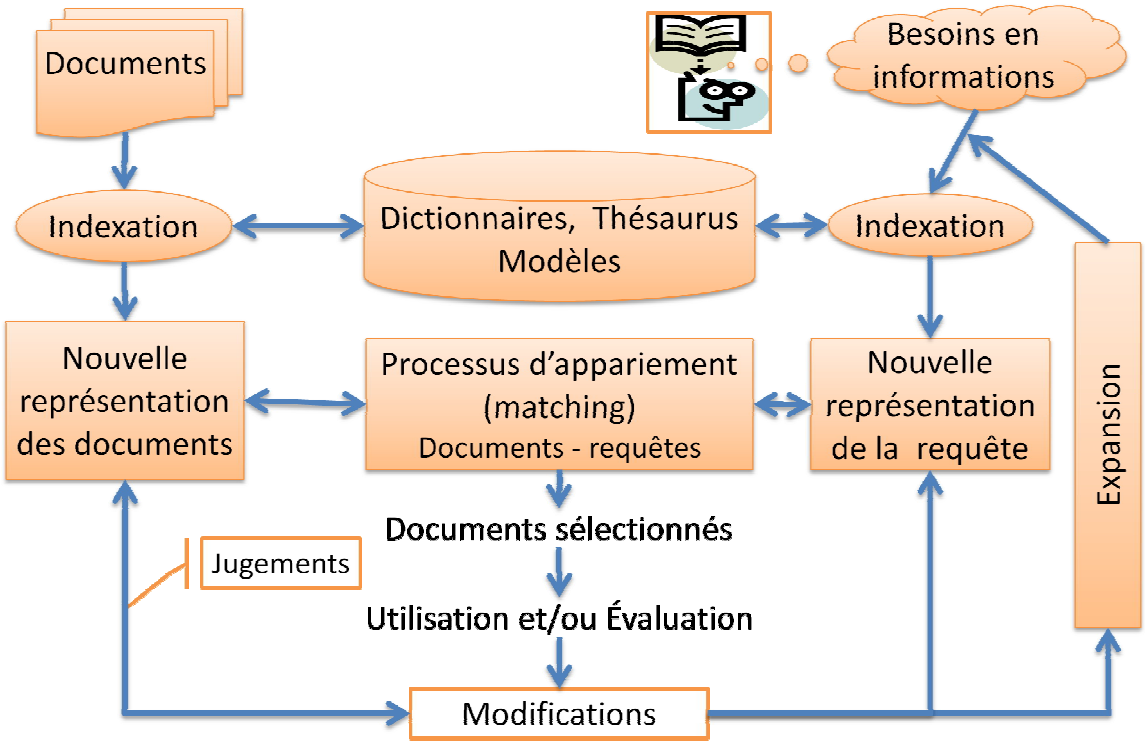

Fig.1.1: Système de recherche d'information.

Le processus de RI a pour but de permettre de retrouver les documents pertinents à une requête de l'utilisateur, à partir d'une base de documents volumineuse. Il s'articule autour de deux étapes essentielles : les phases d'indexation et de recherche. Le processus complet est représenté en figure 1.1.

L'étape d'indexation permet de réaliser le passage d'un document textuel (ou une requête) à une représentation exploitable par un modèle de RI par la construction de mots clés appelé Index.

Document textuel (ou requête) $\xrightarrow{\text{Indixation}}$ représentation exploitable par un modèle RI

Alors, l'***indexation*** consiste à extraire du texte un ensemble de mots clés appelés descripteurs, ces descripteurs vont représenter le document dans le corpus. Chaque descripteur peut être accompagné de connaissances à priori pour mieux appréhender la recherche.

La phase de ***recherche*** a pour objectif d'***apparier*** les documents et la requête de l'utilisateur en comparant leurs descripteurs respectifs. Elle se base sur un formalisme précis défini par un modèle de RI. Les documents présentés en résultat à l'utilisateur sont considérés comme les plus pertinents.

Certains systèmes de RI permettent l'interaction avec l'utilisateur, afin d'améliorer la recherche. L'utilisateur intervient à chaque étape pour assister le système à sélectionner les documents qu'il ***juge*** pertinents pour sa requête. Cette interaction peut aussi servir pour améliorer globalement le fonctionnement du système de RI. Les systèmes RI peuvent comporter des composantes supplémentaires pour ***reformuler*** automatiquement la requête.

La pertinence d'un document pour une question posée par l'utilisateur s'exprime dans les modèles de RI sous la forme d'une valeur de plausibilité déterminée grâce à une heuristique. A titre d'exemple, pour la question « recherche d'information » la réponse sera d'autant plus pertinente si «recherche» et «information » se trouvent localisés dans une même phrase. Il le sera d'autant moins si ces deux mots sont répartis indifféremment dans le texte.

2.1. Modèles de RI

Un modèle de recherche d'information a pour rôle de fournir une formalisation du processus de recherche d'information. Il doit accomplir plusieurs taches dont la plus importante est de fournir un cadre théorique pour la modélisation de la mesure de pertinence. Les objectifs de cette section sont : en premier lieu, nous voulons préparer la scène pour résoudre les problèmes de la recherche d'information que nous essayons de traiter dans cette étude. En second lieu, nous voulons fournir un aperçu rapide des principaux modèles de recherche d'information.

2.1.1. Le modèle booléen

Dans le modèle booléen, un document est représenté comme une conjonction logique de termes (non pondérés), par exemple :

$$d = t_1 \Lambda t_2 \Lambda \ldots \Lambda t_n$$

Où d : document, t_i : les termes de ce document.

Une requête est une expression logique quelconque de termes. On peut utiliser les opérateurs ET (V), OU (Λ) et NON (¬). Par exemple:

$$q = (t_1 \Lambda t_2) \, V \, (t_3 \Lambda \neg t_4)$$

La relation de pertinence $R(d,q)$ entre une requête q et un document d est déterminée par les formules suivantes :

$$R(d, q_1) = \begin{cases} 1 \; si \; q_1 \in d \; ; \; q_i \; est \; un \; terme \; de \; q \\ 0 \; si \; non \end{cases} \qquad (1.1)$$

$$R(d, q_1 \Lambda q_2) = \begin{cases} 1 \; si \; R(d, q_1) = 1 \; et \; R(d, q_2) = 1 \\ 0 \; si \; non \end{cases} \qquad (1.2)$$

$$R(d, q_1 V q_2) = \begin{cases} 1 \; si \; R(d, q_1) = 1 \; et \; R(d, q_2) = 1 \\ 0 \; si \; non \end{cases} \qquad (1.3)$$

$$R(d, \neg q_i) = \begin{cases} 1 \; si \; R(d, q_i) = 0 \\ 0 \; si \; non \end{cases} \qquad (1.4)$$

2.1.1.1. Avantages

Le modèle booléen a les avantages suivants :

- Il est facile à implémenter et son calcul est efficace [Fra92].
- Il permet aux utilisateurs d'exprimer des contraintes structurelles et conceptuelles pour décrire les caractéristiques linguistiques importantes [Mar91]. Les utilisateurs constatent que des caractéristiques de synonymes (reflétés par les clauses OU) et les expressions (représentées par des relations de proximité) sont utiles dans la formulation des requêtes [Coo88].
- Le modèle booléen possède une clarté et une grande puissance expressive. La recherche booléenne est très pertinente si une requête exige une sélection approfondie et non ambiguë.
- la méthode booléenne offre plusieurs techniques pour élargir ou rétrécir une requête.
- Le modèle booléen peut être particulièrement pertinent dans les phases avancées du processus de recherche, en raison de la clarté et de la précision avec lesquelles des relations entre les concepts peuvent être représentés.

2.1.1.2. Inconvénients

Le modèle booléen standard souffre cependant des lacunes suivantes :

- Les utilisateurs éprouvent différentes difficultés pour construire des requêtes booléennes pertinentes pour plusieurs raisons [Fox88]. Parmi lesquelles, ils utilisent les termes de langue naturelle ET, OU, ou NON qui ont une signification différente lorsqu'ils sont utilisés dans une requête.
- Une des erreurs communes commise par des utilisateurs est de substituer l'opérateur logique ET par l'opérateur logique OU lors de la traduction d'une phrase à une requête booléenne. En outre, pour former des requêtes complexes, les utilisateurs doivent se familiariser avec les règles de la priorité et de l'utilisation des parenthèses. Les utilisateurs débutants trouvent une difficulté d'utilisation des parenthèses, particulièrement les parenthèses emboîtées. Enfin, ils sont accablés par la multiplicité des moyens par lesquels une requête peut être structurée ou modifiée, en raison de l'explosion combinatoire des requêtes faisables quant le nombre de concepts augmente. En outre, les utilisateurs ont du mal à identifier et appliquer les différentes stratégies qui sont disponibles pour rétrécir ou élargir une requête booléenne [Lan93] [Mar91].
- Ce modèle ne retourne que seulement les documents qui satisfont exactement à une requête. D'une part, l'opérateur logique ET est trop sévère parce qu'il ne distingue pas entre le cas où aucun des concepts n'est satisfait et le cas où tous (sauf un) sont satisfaisants. Par conséquent, quand plus de trois critères sont combinés avec l'opérateur booléen ET, aucun ou peu de documents sont retrouvés (le problème de Null Output). D'autre part, l'opérateur

booléen OU ne reflète pas combien de concepts ont été satisfaits. Ainsi, souvent trop de documents sont retrouvés (le problème de Output Overload).

- Il est difficile de contrôler le nombre de documents recherchés. Des utilisateurs sont souvent confrontés aux problèmes Null Output ou Output Overload et ils sont perturbés de la façon de modifier la requête pour récupérer un nombre raisonnable de documents.
- L'approche booléenne traditionnelle ne fournit pas un classement de pertinence des documents recherchés, bien que les approches booléennes modernes puissent utiliser quelques astuces pour les classer [Mar91].
- Il ne représente pas le degré d'incertitude ou d'erreur due au problème de vocabulaire [Nic92].

2.1.2. Le modèle probabiliste

Le modèle probabiliste aborde le problème de la recherche d'information dans un cadre probabiliste. Il a été proposé au début des années 1960. Il est basé sur le principe de rang de probabilité, qui déclare qu'un système de recherche d'information est censé classer les documents basés sur leur probabilité de pertinence à la requête [Nic92]. Le principe tient compte qu'il y a une incertitude dans la représentation des besoins d'informations et des documents.

Soient P et NP représentant respectivement la pertinence et la non-pertinence des documents pour une requête donnée, le modèle probabiliste tente de déterminer les probabilités P(P|D) et P(NP|D). Ces deux probabilités signifient que : si on retrouve le document D, quelle est la probabilité pour que l'information soit pertinente ou non ?

Dans un premier temps, travaillons sur le contexte suivant :

On considère que la présence (la valeur 1) et l'absence (la valeur 0) de termes dans les documents et dans les requêtes comme des caractéristiques observables.

On suppose qu'on a une requête fixe. On tente de déterminer les caractéristiques de P et NP pour cette requête.

Donc, implicitement, P(P|D) et P(NP|D) correspondent plutôt à $P(P_R|D)$ et $P(NP_R|D)$ pour la requête R, mais cet index peut être ignoré pour l'instant.

Si on peut calculer ces deux probabilités, alors on pourra classer les documents selon ces deux probabilités ou selon la fonction (O(D)) suivante qui compare les deux probabilités :

$$O(D) = P(P|D) / P(NP|D) \qquad (1.5)$$

Plus O(D) est élevée pour un document, plus ce document doit être classé à un rang supérieur.

Cependant, les deux probabilités nécessaires ne sont pas directement calculables. Ainsi, on utilise le théorème de Bayes:

$$P(P|D) = P(D|P)\ P(P) / P(D) \qquad (1.6)$$

$$P(NP|D) = P(D|NP)\ P(NP) / P(D) \qquad (1.7)$$

Où P(D|P) = la probabilité que D fait partie de l'ensemble pertinent,

P(P) = la probabilité de pertinence, c'est-à-dire, si on choisit un document au hasard dans le corpus, la chance de tomber sur un document pertinent ;

P(D) = la probabilité que le document soit choisi (si on prend au hasard un document dans le corpus, la chance de tomber sur D).

Appliquons dans O(D), nous avons :

$$O(D) = P(P|D) / P(NP|D) = [P(D|P)\ P(P)] / [P(D|NP)\ P(NP)] \qquad (1.8)$$

Comme pour la même requête, P(P) et P(NP) sont des constantes, nous pouvons ré-exprimer O(D) comme suit :

$$O(D) \simeq P(D|P) / P(D|NP) \qquad (1.9)$$

(O(D) est proportionnelle à P(D|P) / P(D|NP)).

Étant donné que l'objectif de la RI est de déterminer le rang des documents, on peut très bien utiliser P(D|P) / P(D|NP) à la place de O(D) exacte. Donc, définissons O(D) comme

P(D|P) / P(D|NP).

2.1.2.1. Avantages

Les approches probabilistes ont les avantages suivants :

- Elles fournissent aux utilisateurs un rang de pertinence des documents recherchés. Par conséquent, elles leur permettent de contrôler le rendement en plaçant un seuil de pertinence ou en spécifiant un certain nombre de documents à afficher.
- il peut être plus facile de formuler les requêtes parce que les utilisateurs ne doivent pas apprendre un langage d'interrogation et peuvent utiliser la langue naturelle.

2.1.2.2. Inconvénients

Cependant, les approches probabilistes ont les inconvénients suivants :

- Ils ont une puissance expressive limitée. Par exemple, l'opération NON ne peut pas être représentée parce que seulement des poids positifs sont utilisés.
- Le modèle probabiliste est limité par l'absence de la structure qui exprime les caractéristiques linguistiques importantes telles que les expressions. Il est également difficile d'exprimer les contraintes de proximité, or cette caractéristique est d'une grande utilité pour les chercheurs expérimentés.
- le calcul des scores de pertinence peut être coûteux.

- Une liste linéaire rangée fournit aux utilisateurs une vue limitée de l'espace d'information et elle ne suggère pas directement comment modifier une requête si elle est nécessaire [Spo93].
- Les requêtes doivent contenir un grand nombre de mots pour améliorer la performance de recherche, et par conséquent les utilisateurs sont confrontés au problème de devoir choisir les mots pertinents qui sont également utilisés dans les documents pertinents.

Si les utilisateurs fournissent au système de recherche un *feedback*, alors cette information est utilisée par les approches statistiques pour recalculer les poids tels que : les poids des termes de la requête dans les documents pertinents sont incrémentés, tandis que les poids des termes de requête qui n'apparaissent pas dans les documents pertinents sont diminués [Sal90]. Il y a plusieurs façons de calculer et de mettre à jour les poids et chacune a ses avantages et ses inconvénients.

2.1.3. Le modèle LSI (Latent Semantic Indexing)

Le modèles Latent Semantic Indexing (LSI) consiste à associer et établir des relations entre les mots clés d'un corpus afin que les systèmes de recherche d'information puissent plus facilement identifier la thématique des documents du corpus.

L'approche consiste à profit de la structure évoluée implicite dans la relation entre les termes et les documents « structure sémantique » afin d'améliorer la détection des documents pertinents sur la base des termes trouvés dans les requêtes.

Diverses techniques de statistique et d'intelligence artificielle en association avec la sémantique du domaine ont été utilisées comme méthodes pour aider à résoudre certains problèmes de recherche. Dans le LSI les liens entre les termes et les documents sont calculés et exploités dans le processus de recherche. L'idée principale du modèle LSI est que les idées dans un texte sont plus reliées aux concepts qu'elles décrivent que les termes de l'index utilisés pour leur description. Ainsi, la correspondance entre un document et une requête donnée devrait être basée sur la correspondance des concepts plutôt que sur la correspondance des termes de l'index. L'objectif fondamental est d'aboutir à une représentation conceptuelle des documents. Ainsi, les requêtes peuvent rechercher des documents même si elles n'ont aucun mot en commun.

Le but de LSI est de transformer une représentation par des mots-clés en une autre représentation qui est « meilleure ». Le mot « meilleure » est compris dans le sens suivant: les documents et les requêtes sémantiquement similaires seront plus proches avec la représentation transformée qu'avec les mots-clés. La transformation par LSI est réalisée comme suit:

1. Au début, chaque document et requête sont représentés comme un vecteur de mots-clés.
2. LSI utilise la SVD[1] pour créer un nouvel espace vectoriel:

$$X = T_0 S_0 D_0' \qquad (1.10)$$

Où :

- X est la matrice document-terme originale (de taille $t \times d$)
- T_0 est une matrice $t \times m$
- S_0 est une matrice $m \times m$ diagonale (seulement les éléments en diagonale sont non-nuls)
- D_0' est une matrice $m \times d$. La valeur m est choisie comme une valeur $\leq \min(t,d)$.

De plus, on trie les valeurs de S_0 dans l'ordre décroissant. Il existe juste une seule décomposition de cette façon.

3. On pense que la représentation par des mots-clés contient beaucoup de bruits. Typiquement, ces bruits se retrouvent dans les dimensions de S_0 qui ont des valeurs faibles. Ainsi, la technique de LSI veut ignorer ces dimensions de valeurs faibles (ou de les ramener à la valeur 0), ce qui ramène les dimensions de S_0 à k, et cette matrice réduite est notée par S. En conséquence, les matrices T_0 et D_0 nettoyées deviennent T et D.
4. Chaque requête soumise, est traduite forcement dans ce nouvel espace.

$$D_q = X_q' T S^{-1} \qquad (1.11)$$

Où X_q est le vecteur de mots-clés de la requête.

Ensuite, ce nouveau "document" est ajouté dans la matrice D. Le calcul de similarité entre chaque paire de documents peut se faire comme suit:

$$D S^2 D' \qquad (1.12)$$

Ainsi, après ce calcul, on peut connaître la similarité de ce nouveau document (ou la requête) avec tous les autres documents.

Ce modèle a montré des performances très intéressantes. Pour un corpus de petite ou moyenne taille, la performance est très supérieure au modèle vectoriel classique, et il est un des meilleurs modèles. Quand la taille du corpus augmente, la différence avec les autres modèles classiques semble diminuer.

---

[1] En mathématiques, le procédé d'algèbre linéaire de décomposition en valeurs singulières (ou SVD, de l'anglais : Singular Value Decomposition) d'une matrice est un outil important de factorisation des matrices rectangulaires réelles ou complexes. Ses applications s'étendent du traitement du signal aux statistiques, en passant par la météorologie.

2.1.4. Le Modèle vectoriel

C'est un autre modèle souvent utilisé. Il représente les documents et les requêtes comme vecteurs de poids dans un espace multidimensionnel, dont les dimensions sont les termes utilisés pour construire un index qui représente les documents [Sal83].

$$D = \begin{bmatrix} \textbf{est} & \textbf{un} & \textbf{autre} & \textbf{modèle} & \textbf{souvent} & \textbf{utilisé} & \textbf{représente} & \textbf{documents} & \textbf{requêtes} & \textbf{comme} & \textbf{vecteurs} & \textbf{poids} \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$$R = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Fig.1.2: Exemple sur la représentation des documents et des requêtes.

La création d'un index implique une lecture lexicologique pour identifier les termes significatifs, où l'analyse morphologique ramène les différentes formes de mot aux « lemmes » communs, et l'occurrence de ces lemmes est calculée. Des substituts de requête et de document sont comparés selon leurs vecteurs. Par exemple, Soit l'espace vectoriel suivant:

$$<t_1, t_2, t_3, \ldots, t_n> \qquad (1.13)$$

Un document et une requête peuvent être représentés comme suit:

$$d = < a_1, a_2, a_3, \ldots, a_n> \qquad (1.14)$$

$$q = < b_1, b_2, b_3, \ldots, b_n> \qquad (1.15)$$

Ainsi, a_i et b_i correspondent aux poids du terme t_i dans le document et dans la requête.

Étant donnés ces deux vecteurs, leur degré de correspondance est déterminé par leur *similarité*. Il y a plusieurs façons de calculer la similarité entre deux vecteurs. En voici quelques unes:

$$Sim0(d, q) = \Sigma_i (a_i * b_i) \quad \text{(produit interne)} \qquad (1.16)$$

$$Sim1(d, q) = \Sigma_i (a_i * b_i) / [\Sigma_i (a_i)^2 * [\Sigma_i (b_i)^2]^{1/2} \quad \text{(cosinus)} \qquad (1.17)$$

$$Sim2(d, q) = 2\ \Sigma_i (a_i * b_i) / [[\Sigma_i (a_i)^2 + [\Sigma_i (b_i)^2] \qquad (1.18)$$

$$Sim3(d, q) = \Sigma_i (a_i * b_i) / [[\Sigma_i (a_i)^2 + [\Sigma_i (b_i)^2 - \Sigma_i (a_i * b_i)] \qquad (1.19)$$

Sauf la première formule, toutes les autres sont normalisées, c'est-à-dire qu'elles donnent une valeur dans [0, 1].

Dans le modèle vectoriel, les termes d'un substitut de requête peuvent être pesés pour tenir compte de leur importance, et ils sont calculés en utilisant les distributions statistiques des termes dans la collection des documents [Sal83]. Ce modèle peut assigner un haut classement à un document qui contient seulement quelques termes de requête si ces termes se produisent rarement dans la collection mais fréquemment dans le document.

2.1.4.1. Avantages

Le modèle vectoriel a les avantages suivants :

- Le langage de requête est plus simple (liste de mot clés).
- Les performances sont meilleures grâce à la pondération des termes.
- Le renvoi de documents à pertinence partielle est possible.
- La fonction d'appariement permet de trier les documents.

2.1.4.2. Inconvénients

Cependant, ce modèle a les inconvénients suivants :

- Le modèle considère que tous les termes sont indépendants.
- Le langage de requête est moins expressif.
- De temps en temps l'utilisateur ne sait pas pourquoi un document est retourné par le système.

2.2. Critères d'évaluation des SRI

La pertinence d'un document pour une requête de l'utilisateur s'exprime dans les modèles de RI sous la forme d'une valeur de plausibilité déterminée grâce à une heuristique. Alors la pertinence est une connaissance très complexe à évaluer. Ainsi, elle dépend fortement de l'utilisateur, de plus les utilisateurs d'un système de RI ont des besoins très variés et des critères assez différents pour juger si un document est pertinent. Il est néanmoins essentiel de disposer de techniques d'évaluation solides qui en définissant des mesures précises, permettent de juger l'efficacité des systèmes RI à retrouver des documents pertinents.

2.2.1. Évaluation

Les systèmes de RI sont toujours évalués en fonction de la pertinence des documents retrouvés. Afin de procéder à des évaluations automatiques, nous avons besoin de corpus de test « standard ». Chaque corpus contient :

L'ensemble de documents ;

L'ensemble de requêtes de test sur l'ensemble de documents du même corpus;

La liste de documents pertinents pour chaque requête.

Un système de RI quelconque peut utiliser ce corpus pour trouver des documents pour les requêtes données, et nous pouvons comparer ces documents retrouvés avec la liste de documents pertinents pour évaluer la qualité du système.

Les deux principales mesures utilisées pour évaluer un système de RI sont la précision et le rappel. Ces mesures reflètent la comparaison des réponses d'un système pour l'ensemble des requêtes avec les réponses idéales (liste de documents pertinents dans le corpus). Plus précisément, ces deux mesures peuvent être définies par :

2.2.1.2.Précision

Un système de RI sera très précis si presque tous les documents retrouvés sont pertinents. En fait c'est la proportion des documents pertinents parmi l'ensemble de ceux retrouvés par le système.

$$précision = \frac{nombre\ total\ de\ documents\ pertinents\ retrouvés\ par\ le\ système}{nombre\ total\ de\ documents\ retrouvés\ par\ le\ système} \qquad (1.20)$$

2.2.1.3.Rappel

Un système de RI aura beaucoup de rappel s'il retrouve la plupart des documents pertinents du corpus pour une requête. En fait c'est la proportion de documents pertinents retrouvés par le système parmi tous ceux qui sont pertinents

$$rappel = \frac{nombre\ total\ de\ documents\ pertinents\ retrouvés\ par\ le\ système}{nombre\ total\ de\ documents\ pertinents\ dans\ le\ corpus} \qquad (1.21)$$

L'idéal pour un système de RI est d'avoir de bons taux de précision et de rappel en même temps. Les deux métriques ne sont pas indépendantes.

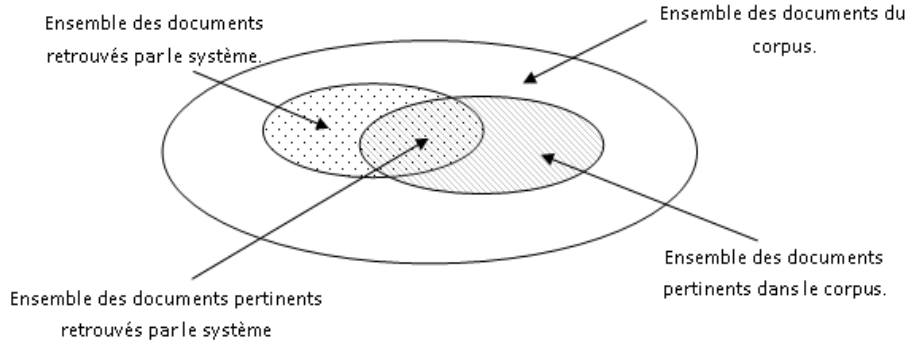

Fig.1.3: Rappel/Précision

2.2.2. La courbe de Rappel/Précision

Les performances d'un système de RI peuvent être représentées par une courbe Rappel/Précision. Lorsque les valeurs exactes de rappel ne peuvent pas être atteintes, Il est fréquent d'employer une interpolation sur ces courbes, qui consiste à lisser la courbe initiale pour qu'elle soit décroissante.

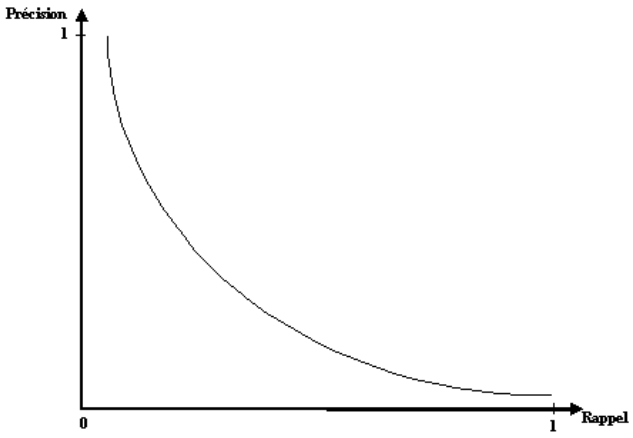

Fig.1.4: Exemple de courbe précision/rappel.

L'idéal pour un système de RI est d'avoir de bons taux de précision et de rappel en même temps. Il y a une forte relation entre elles : quand l'une augmente l'autre diminue. En pratique, la précision évolue en fonction du rappel et vice versa.

2.2.3. Mesures globales

2.2.3.1 La précision moyenne interpolée IAP (Interpolated Average Precision)

Est une mesure décrivant la précision globale du système évalué sur une requête. Elle consiste simplement à faire la moyenne des 11 précisions interpolées obtenues pour les points de rappels fixes, de 0 %, à 100 % par pas de 10 %. La règle d'interpolation est la suivante : la valeur interpolée de la précision pour un niveau de rappel *i* est la précision maximale obtenue pour tous les rappels supérieurs ou égaux à *i*.

2.2.3.2. La R-précision

La R-précision est calculée pour une requête, la précision obtenue pour un nombre donné de documents retrouvés par le système. Ce nombre est fixé pour chaque requête en fonction du nombre de documents pertinents dans la collection. La R-précision est intéressante lorsque la collection comporte un nombre considérable de documents pertinents.

2.2.3.3. La F-mesure [Van79]

La F-mesure prend en considération la précision et le rappel simultanément. Elle est définie comme la combinaison pondérée du taux de rappel et du taux de précision:

$$F = \frac{2P * R}{P + R} \qquad (1.22)$$

Où P et R représentent respectivement les résultats de précision et de rappel.

3. RI en langue arabe

L'arabe, une des six langues officielles des Nations Unies, est la langue maternelle de plus de 300 millions de personnes[1]. Le domaine RI arabe est devenu un centre de la recherche et du développement commercial du à la nécessité essentielle de tels outils pour des personnes dans l'ère électronique. Le nombre d'internautes arabophones en 2002 était environ 4.4 millions, environ 1.5% de la population du monde arabe[2]. Mais, de l'autre côté de la réalité, peu de moteurs de recherche sont mis à la disposition des utilisateurs arabophones, bien que les efforts soient en marche pour servir le nombre croissant d'utilisateurs. Récemment des normes pour l'évaluation des outils disponibles ont été introduites par TREC et CLEF en 2000.

Une étude rapide des outils de RI arabes montre deux catégories importantes [Abd04] :

- **Le RI sur la base pleine forme:** La plupart des outils d'indexation et de recherche commerciales utilisées sont des systèmes sur la base pleine forme c'est à dire l'unité de recherche est le mot complet. Ces systèmes incluent le moteur de Web de Sakhr www.alidrisi.com et www.ayna.com et d'autres moteurs multilingues à l'instar de www.google.com ou www.alltheweb.com.
- **Le RI sur la base Morphologie*:*** Les efforts qui ont été faits dans l'environnement académique pour évaluer des systèmes plus sophistiqués donnent une idée au sujet de la nouvelle génération des systèmes d'indexation et de recherche arabe. L'évaluation a été exécutée sur des systèmes utilisant différentes approches de la morphologie (lemme, racine, lemme légère [Lar02b] [Gey01]) et autres basés sur la non-règle; des lemmatiseurs statistiques ou de modèles n-gramme ont été testés visant cet objectif. Généralement, l'utilisation des lemmatiseurs améliore le rappel et la précision. [Lar02b] les expériences ont prouvé que le lemmatiseur léger est plus performant que le lemmatiseur régulier.

3.1. Les ressources arabes :

Un des obstacles principaux pour les développeurs des systèmes de RI arabe est le manque de ressources adéquates qui pourraient aider à tester leurs systèmes afin d'obtenir leur bonne évaluation dans le monde réel.

3.1.1. Corpus

La collection de TREC, une parmi ces ressources de large échelle connue et disponible pour les utilisateurs, elle représente un volume de 884 MOctets. Ce corpus est constitué de 383 872 documents. Il a été encodé en utilisant le SGML et a été transcodé à Unicode (UTF-8). Il inclut des articles journalistiques provenant d'Arabic Newswire de l'AFP (Agence France Presse) du 13 mai 1994 au 20 décembre 2000 avec approximativement 76 millions d'unités lexicales.

[1] http://www.frcu.eun.eg/www/homepage/cdc/cdc.htm. Accédé le 12/12/2008.
[2] http://sonomabusiness.com/archives/2002-09-column-levini.html. Accédé le 05/12/2008.

	TREC2001	TREC2002
Langue du corpus des documents	Arabe	arabe
Nombre de documents	383 872	383 872
Capacité du corpus (MB)	884	884
Nombre total de mots (tokens)	76 millions	76 millions
Nombre de mots différents	666 094	666 094
Taille moyenne des documents (mots)	150	150
Langues des requêtes	arabe et anglais	arabe et anglais
Nombre de requêtes	25	50
Taille moyenne des requêtes (mots)	12.6 (arabe) 12.1 (anglais)	11.2 (arabe) 11 (anglais)
Nombre moyen de documents pertinents par requête	164	118.18

Tableau 1.1: Caractéristiques de la collection TREC arabe.

3.1.2. Dictionnaire

D'autres ressources telles que les dictionnaires monolingues et les dictionnaires bilingues sont nécessaires; ces types de ressources peuvent varier des dictionnaires de traduction automatique aux dictionnaires manuels pour un sujet ou une utilisation spécifique. Les dictionnaires Ajeeb et Ectaco sont accessible en ligne, ils ont été utilisés dans quelques expériences de RI [Lar03]. De plus d'autres efforts ont été déployés dans différentes applications [Zaj01].

3.1.3. Outils

3.1.3.1. Analyseurs morphologiques

L'analyseur morphologique segmente les unités lexicales, repère les différents composants et atteste leur appartenance à la langue.

3.1.3.2. Les concordanciers

Le concordancier a pour objectif de permettre l'exploration du corpus selon les traits proposés par l'analyse morphologique et selon les informations graphiques qui se trouvent dans le texte. Il prend en entrée un texte ou un ensemble de textes et il permet :

- La construction de listes de fréquences d'items, de racines ou tout autre trait de l'analyse morphosyntaxique, par ordre alphabétique ou par ordre fréquentiel.
- La construction d'une concordance.

3.1.3.3. Racineurs

Le racineur consiste à détecter la racine d'une unité lexicale, Les algorithmes de racinisation en arabe les plus connus sont ceux de larkey [Lar05] et Khoja [Kho99].

3.2. Lemmatisation

L'objectif de la lemmatisation est de trouver la forme représentative d'index d'un mot à partir de sa forme représentée dans le document par l'application de la troncature des affixes. La question qui se pose pour la lemmatisation est la suivante : quel lemme linguistique doit-on choisir à un mot pour fin de recherche d'information ? Il ne sera pas suffisant pour la RI arabe de tronquer seulement un préfixe et seulement un suffixe de ce mot.

Dans la langue Arabe, le traitement morphologique devient particulièrement important pour la recherche d'information, parce que la RI vise à déterminer une forme appropriée d'index aux mots. Un des traitements les plus importants pour la langue arabe, en vue de la recherche d'information, est la **lemmatisation** des mots. Cependant, cette langue renferme un certain nombre de complexités (voir chapitre 4) qui rend son traitement difficile.

4. Conclusion

Le but de ce chapitre était de présenter l'état de l'art du domaine de la recherche d'information, de décrire plus particulièrement les principales étapes à savoir l'indexation et la recherche.

Le modèle joue un rôle central dans la RI. C'est le modèle qui détermine le comportement clé d'un système de RI. Dans ce chapitre nous avons discuté des quatre approches principales de modélisation de la recherche d'information, à savoir le modèle booléen, le modèle probabiliste, le modèle LSI (Latent Semantic Indexing), et le modèle vectoriel. D'autre part, Nous avons défini les notions fondamentales de la discipline comme la pertinence des documents par rapport à une requête, l'évaluation des systèmes et les principales méthodes d'évaluation utilisées pour estimer les performances d'une technique par rapport à l'autre.

La langue arabe est l'une des langues les plus largement répandues dans le monde, pourtant il y a relativement peu d'études sur la restitution par des moteurs de recherche de documents pertinents en arabe. Dans ce cadre nous avons présenté quelques ressources linguistiques de l'arabe ainsi que les outils qui sont bien connus dans le domaine de la recherche d'information.

Dans le chapitre qui suit, nous allons centrer notre intérêt sur les propriétés morphologiques de la langue l'arabe.

CHAPITRE 2 : PROPRIETES MORPHOLOGIQUES DE L'ARABE

1. Introduction:

La langue arabe est d'une origine très différente des langues européennes. Elle fait partie du groupe des langues sémitiques. Ce groupe se divise en langues sémitiques orientales, sémitiques occidentales et sémitiques méridionales. À la différence d'autres nations; telles que les anciens égyptiens, les babyloniens et les chinois dont les systèmes d'écriture remontent à des milliers d'années, l'écriture arabe n'est apparue qu'au VI^{e} siècle.

Certains historiens et chercheurs pensent que l'origine de l'écriture arabe est le syriaque en se basant sur:

- l'ordre primitif des lettres arabes;
- les formes de l'ancien alphabet arabe dit « coufique », qui sont comparables à celles de « l'estranghelo »; une forme de l'écriture syriaque.

L'arrivée de l'Islam a profondément marqué l'histoire de la langue et de l'écriture arabe. Le Coran, livre sacré recueillant la parole de Dieu, mais aussi code juridique et moral, occupe d'emblée une place centrale dans la vie du croyant et de la communauté musulmane.

Les recherches sur le traitement automatique de l'arabe ont débuté vers les années 1970, et les premiers travaux concernaient notamment les lexiques et la morphologie.

Ainsi, la croissance rapide du World Wide Web (WWW) et l'explosion primordiale des ressources de l'indexation et des outils de recherche, influent effectivement sur le développement dans le domaine de la recherche d'information multilingue.

Les statistiques prouvent que depuis 1995 quand le premier journal arabe fut lancé en ligne sur www.asharqalawsat.com, le nombre de sites Web arabes s'était développé exponentiellement. Et dès l'année 2000, plus de 20 mille sites arabes sont introduits sur le Web, dont plus de 7% édités [Abdb04].

2. Particularité de la langue arabe

Il est évident que les 22 consonnes que compte l'alphabet arabe sont insuffisantes pour écrire les 28 phonèmes arabes. À ce problème s'ajoute l'absence de points diacritiques et les signes vocaliques qui réduisent la graphie à un simple schéma consonantique. La réforme de l'écriture arabe devenait alors une affaire impérative et très urgente.

Les philologues arabes ont inventé six nouvelles lettres. Il s'agit des lettres suivantes : ث, خ, ذ, ض, ظ, غ.

Des petits points noirs ont été utilisés comme marques de différenciation entre des lettres qui partageaient une forme identique. Ces points sont placés au-dessus et au-dessous de la lettre en un, deux ou trois.

Exemples :

ب ,ت , ث ;

ف , ق ;

خ , ج , etc.

L'arabe s'écrit et se lit cursivement (écriture dont les lettres sont reliées les unes avec les autres) de droite à gauche, en utilisant un alphabet de 28 lettres (Tableau 2.1).

Lettre arabe	Prononciation	Lettre arabe	Prononciation
أ	Alif	ض	Dad
ب	Ba'	ط	Tah
ت	Ta'	ظ	Zah
ث	Tha'	ع	Ayn
ج	Jim	غ	Ghayn
ح	Hha'	ف	Fa
خ	Kha'	ق	Qaf
د	Dal	ك	Kaf
ذ	Thal	ل	Lam
ر	Ra	م	Mim
ز	Zayn	ن	Nun
س	Sin	ه	Ha
ش	Chin	و	Waw
ص	Sad	ي	Ya

Tableau 2.1: Les 28 lettres arabes.

La représentation morphologique de l'arabe est assez complexe en raison de la variation morphologique et du phénomène d'agglutinement; les lettres changent de formes selon leur position dans le mot (isolée, initiale, médiane et finale). Le tableau 2.2 montre un exemple des différentes formes de la lettre «qaf» dans différentes positions. Nous pouvons observer ainsi plusieurs caractéristiques générales de cette langue suivant le détail ci-après :

Isolée	Initiale	Médiane	Finale
ق	قـ	ـقـ	ـق
	قـرآن	الـقـرآن	غسـق

Tableau 2.2: différentes écritures de la lettre « qaf » en différentes positions dans le mot et comme lettre séparée.

Pour une meilleure précision de la prononciation, des signes ont été inventés. Il s'agit de trois voyelles brèves et de sept signes orthographiques qui s'ajoutent aux consonnes.

Ces trois voyelles brèves sont :

- Fatha «ـَـ», elle surmonte la consonne et se prononce comme un «a» français ;
- Damma «ـُـ», elle surmonte la consonne et se prononce comme un «ou» français ;
- Kasra «ـِـ», elle se note au-dessous de la consonne et se prononce comme un « i » français).

  Les sept signes orthographiques sont :

- Sukun «ـْـ» : ce signe indique qu'une consonne n'est pas suivie (ou muet) par une voyelle. Il est noté toujours au-dessus de la consonne;
- Les trois signes de tanwin : lorsque (la Fatha, la Kasra et la Damma) sont doublées, elles prennent un son nasal, comme si elles étaient suivies de «n» et on les prononce respectivement :
  - an «ـًـ» pour les Fathatan ;
  - in «ـٍـ» pour les Kasratan ;
  - un «ـٌـ» pour les Dammatan.
- Chadda «ـّـ» comme dans le français, l'arabe peut renforcer une consonne quelconque;
- Wasla « ٱ » : quand la voyelle d'un Alif au commencement d'un mot doit être absorbée par la dernière voyelle du mot qui précède;
- Madda « آ » : la madda (prolongation) se place sur l'Alif pour indiquer que cette lettre tient lieu de deux alifs consécutifs ou qu'elle ne doit pas porter le Hamza.

Le tableau 2.3 montre un exemple pour les mots «كتب » et «شعر».

	Interprétation I		Interprétation II		Interprétation III	
كتب	كَتَبَ	Il a écrit	كُتِبَ	Il a été écrit	كُتُبٌ	Des livres
شعر	شَعَرَ	Il a senti	شِعْرٌ	Poème	شَعْرٌ	Chevelure

Tableau 2.3: ambiguïté causée par l'absence de voyelles pour les mots «كتب » et «شعر».

Cependant, les textes courants rencontrés dans les journaux et les livres ne comportent habituellement pas de voyelles. De plus, certaines lettres comme Alif «ا» peuvent symboliser le «أ», «إ», «آ» ou «ٱ»; de même que pour les lettres «ى» et «ه» qui symbolisent respectivement «ي» et «ة» [Xuj02].

3. Morphologie arabe

Le lexique arabe comprend trois catégories de mots : verbes, noms et particules. La plupart de noms et de verbes sont dérivés d'un nombre réduit (approximativement 10000) de racines. Ces racines sont les unités linguistiques portant une signification sémantique, et la plupart d'entre elles se composent seulement de 3 consonnes, et rarement de 4 ou de 5 consonnes.

De ces racines, nous pouvons produire des dérivés nominaux et verbaux par l'application des modèles (règles morphologiques). On peut produire jusqu'à 30 mots d'une racine de 3 consonnes; le tableau 2.4 montre quelques exemples de schèmes appliqués aux mots «كتب» écrire et «حمل» porter, à partir desquels nous pouvons produire plusieurs mots :

Modèles	كتب	KTB	Notion d'écrire
ـاعِل	ـَاتِب	Katib	Écrivain
ـَعَل	تَبَ	Katab	a écrit
فْعَل	كْتَب	Makta	bureau
ـعِل	تِبَ	Kutib	A été écrit
...			

حمل	HML	Notion de porter
ـَامِل	Hamil	porteur
مَلَ	Hamala	a porté
حْمَل	mahmal	brancard
مِلَ	Humila	a été porté

Tableau 2.4: Exemple de modèles pour les mots «كتب » écrire et «حمل » porter.

L'Arabe comprend environ 150 modèles (schémas ou patrons) dont certains plus complexes, comme le redoublement d'une consonne ou l'allongement d'une voyelle de la racine, l'adjonction d'un ou de plusieurs éléments ou encore la combinaison des deux.

4. Structure d'un mot

La définition du mot du point de vue du traitement automatique se heurte à des considérations syntaxiques et sémantiques. Dans le domaine des langages formels, la transformation du flux de caractères représentant un texte en une suite d'unités mieux adaptées aux traitements ultérieurs, est habituellement appelée segmentation (tokenization), et les unités produites les segments (tokens) sont construites sur la base de définitions purement orthographiques. Le problème posé par de telles techniques pour des applications de traitements de langue est malheureusement l'absence de correspondance biunivoque entre les segments ainsi identifiés et les unités textuelles élémentaires (les mots) manipulées dans le traitement linguistique. En arabe cette séquence de lettres est appelée le mot graphique.

Les mots sont séparés par des espaces et d'autres signes de ponctuation. Néanmoins, des prépositions sont agglutinées au mot (apparaissant après eux), faisant des limites invisibles entre le mot et la préposition.

Plusieurs types d'affixes sont agglutinés au début et à la fin des mots : antéfixes, préfixes, suffixes et post fixes.

Antéfixe	Préfixe	Noyau	Suffixe	Post fixe

Tableau 2.5: Structure d'un mot arabe.

Ainsi nous pouvons les classer par catégorie selon leur rôle syntaxique.

4.1. Les antéfixes:

Les antéfixes sont généralement des prépositions agglutinées au début des mots. Ils se combinent entre eux pour donner les traits syntaxiques, coordonnant, terminant ...etc.

Voici une liste non exhaustive des antéfixes simples.

- La coordination par les coordonnants « فَ» fa et « وَ wa.
- L'interrogation par le morphème « أَ a.
- La marque du futur « سَ » sa.
- L'article « اَلْ» al.
- Les prépositions par les lettres « بِ» bi et « لِ li.
- Les particules du subjonctifs النصب « فَ» fa, « لِ li, et « وَ wa.
- Le marqueur de comparaison par les lettres « كَ» ka.
- Le marqueur de corroboration « لَ» la.
- La particule du jussif (الجزم) par la lettre « لِ li.

4.2. Les préfixes:

Les préfixes, habituellement représentés par une seule lettre, indiquent la personne de conjugaison des verbes au présent. Ils ne se combinent pas entre eux.

Numéro de préfixe	Préfixe
1	
2	
3	تَ
4	نُ
5	نَ
6	نُ
7	يَ
8	يُ

Tableau 2.6: listes des préfixes arabes.

4.3. Les suffixes:

Les suffixes sont les terminaisons de conjugaison des verbes et de marques duelles/plurielles/femelles pour les noms y compris les adverbaux. Ils ne se combinent pas entre eux.

Voici la liste exhaustive de tous les suffixes:

N°Suff	Suffixe
1	◌ً
2	
3	◌ٌ
4	◌ٍ
5	◌َ
6	◌َا
7	◌َاتٌ
8	◌َاتٍ
9	◌َاتُ
10	◌َاتِ
11	انِ
12	◌َانٍّ
13	◌َةً
14	◌َةٌ
15	◌َةٍ
16	◌َةَ
17	◌َةُ

N°Suff	Suffixe
18	◌َةِ
19	◌َتَا
20	تَانِ
21	◌َتَيْ
22	تَيْنِ
23	◌َتْ
24	◌َنَّ
25	◌َنْ
26	◌َوْا
27	◌َوْنَ
28	◌َيْ
29	◌َيْنَ
30	يْنِ
31	◌ُ
32	◌ُنَّ
33	◌ُنْ
34	◌ُو

N°Suff	Suffixe
35	◌ُوا
36	◌ُونَ
37	◌ِ
38	◌ِنَّ
39	◌ِنْ
40	◌ِي
41	◌ِينَ
42	◌ْ
43	◌ْتَ
44	◌ْتُ
45	تُمَا
46	◌ْتُمْ
47	◌ْتُنَّ
48	◌ْتِ
49	◌ْنَ
50	◌ْنَا
51	◌ْنَانِّ

N°Suff	Suffixe
52	تَ
53	تُ
54	مَا
55	تُمْ
56	تُنَّ
57	تِ
58	نَ
59	نَا
60	نَانِّ
61	ِيَّا
62	◌ِيٌّ
63	◌ِيٍّ
64	◌ِيَّ
65	◌ِيُّ
66	◌ِيِّ

Tableau 2.7: listes des suffixes arabes.

4.4. Les post fixes:

Finalement, les post fixes représentent des pronoms attachés à la fin des mots. Ils peuvent se combiner entre eux.

Voici dans le tableau suivant une liste des post fixes:

N° post fixe	post fixe	Description
1	ـِي	1er Personne, Masculin/Féminin, Singulier
2	ـي	1er Personne, Masculin/Féminin, Singulier
3	ـنَا	1er Personne, Masculin/Féminin, Duel/Pluriel
4	ـكَ	2eme Personne, Masculin, Singulier
5	ـكِ	2eme Personne, Féminin, Singulier
6	ـكُمَا	2eme Personne, Masculin/Féminin, Duel
7	ـكُمْ	2eme Personne, Masculin, Pluriel
8	ـكُنَّ	2eme Personne, Féminin, Pluriel
9	ـهُ	3eme Personne, Masculin, Singulier
10	هَا	3eme Personne, Féminin, Singulier
11	هُمَا	3eme Personne, Masculin/Féminin, Duel
12	هُمْ	3eme Personne, Masculin, Pluriel
13	هُنَّ	3eme Personne, Féminin, Pluriel
14	ـهِ	3eme Personne, Masculin, Singulier
15	ـهِمَا	3eme Personne, Masculin/Féminin, Duel
16	ـهِمْ	3eme Personne, Masculin, Pluriel
17	ـهِنَّ	3eme Personne, Féminin, Pluriel

Tableau 2.8: listes des post fixes arabes.

Tous ces affixes devraient être traités correctement pendant la lemmatisation de mots.

Exemple

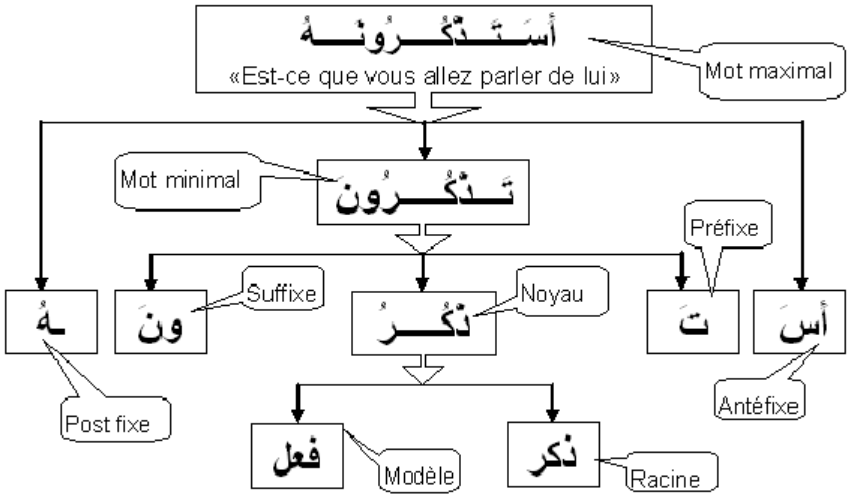

Fig.2.1 : Segmentation du mot en arabe« أَسَتَذْكُرُونَهُ»«Est-ce que vous allez parler de lui»

5. Les catégories des mots

L'arabe comprend trois catégories de mots : verbe, nom et particule**.**

5.1. Verbe

Nous pouvons classer les verbes arabes selon plusieurs critères : Selon le nombre et la nature des consonnes de leurs racines, et selon leurs modèles. En classant les verbes selon le nombre des consonnes de la racine, nous aurons soit des verbes trilitères qui ont trois consonnes, soit des verbes quadrilatères, peux nombreux, qui ont quatre consonnes.

Selon le modèle et le nombre de consonnes qui constituent la structure verbale, nous avons soit des verbes nus (مُجرّد) qui sont composés seulement par les consonnes de leurs racines et des voyelles brèves, soit des verbes augmentés ou dérivés (مزيد) qui sont dérivés de trois consonnes de la racine par modification des voyelles, par redoublement de la deuxième lettre de la racine, par adjonction et même par intercalation d'affixes.

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).

5.2. Nom

Les noms arabes regroupent les substantifs, les adjectifs et les pronoms, ainsi que d'autres noms invariables. Les substantifs et les adjectifs sont créés en prenant pour origine tantôt un type verbal, tantôt un type nominal. Nous pouvons distinguer dans le tableau 2.9 deux classes de

noms : la première regroupe les noms conjugables ou semi-conjugables qui peuvent avoir la forme duelle, plurielle, etc. la deuxième classe regroupe les noms non-conjugables qui gardent la même forme quelque soit le contexte. Les noms conjugables sont soit des noms primitifs, qui échappent à toute dérivation comme «غَزَالٌ (gazelle), soit des noms dérivationnels qui sont formés à partir d'une racine comme « مَكْتَبَةٌ (bibliothèque) de la racine «كتب».

5.3. Particule

Les particules sont des lemmes invariables et en nombre limité. Ils indiquent l'articulation de la phrase. Elles sont classées selon leur champ sémantique et leur fonction dans la phrase; on en distingue plusieurs types :

- Préposition : exemple (حتَّى، عَنْ ، ل، ك، ب)
- Particules de coordination : exemple (و،ف، ثُمَّ، أوْ)
- Particules interrogatives : exemple (مَا، هلْ، أ)
- Particules d'affirmation : exemple (نَعَمْ، بَلى، أَجَلْ)
- Particules de négation : exemple (لا، لَنْ ، لَمْ)
- Particules distinctive : exemple (أيْ)
- Particules relatives : exemple (مَا)
- Particules de futur : exemple (سَوْفَ ، سَ)
- Particules conditionnelles : exemple (إنْ، لَوْ)

Ces particules seront très utiles pour notre traitement, elles font partie du dictionnaire qui regroupe les mots vides.

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

Catégorie	Dérivation	Conjugaison	Sous-catégorie	Exemples
Nom	Dérivationnel irrégulier	Non conjugable	Adverbe	أَيْنَ ، حَيْثُ
			Nom de voix	، نِحْ
			Nom de verbe	هَيْهَاتَ ، آهْ، أُفٍّ
			Pronom Personnel (affixé ou isolé)	هُوَ ، أَنَا تُ ، نَ
			Pronom interrogatif	ـَ ، مَتَى، مَا
			Pronom conditionnel	نْ ، إِذَا
			Pronom allusif	، كَأَيّ
		Conjugable	Pronom relatif	الذي، التي
			Nom de nombre	، خَمْسَةٌ
			Pronom démonstratif	ا، هَذِهِ
			Nom propre	مَّدٌ، زَيْنَبٌ
			Nom commun	لٌ ، قَلَمٌ
	Dérivationnel régulier	Conjugable	Masdar	ابَةٌ ، الحَيَاةُ
			Participe actif	ِلٌ ، صَائِمٌ
			Participe passif	وجُودٌ، كْتُوبٌ
			Nom d'une fois	جَلْسَةٌ ، حَجْةٌ
			Nom de manière	َةٌ ، نِظْرَةٌ
			Nom de temps	ْرِبٌ
			Nom de lieu	بٌ ، مَقْبَرَةٌ
			Nom d'instrument	ارٌ ، مِطْرَقَةٌ
			Adjectif	نٌ ، فَحْلٌ ،
			Elatif	َنُ ، أَفْضَلُ
			Nom diminutif	كُتَيِّبٌ ، شُوَيِّعِرٌ
			Nom de relation	ِرِيٌّ ، لُبْنَانِيٌّ
			Intensif	احٌ ، غَوَّاصٌ

Tableau 2.9: Classement des sous catégories de noms.

6. Problèmes du traitement automatique de l'arabe

En Arabe écrit, les voyelles (signes diacritiques) sont omises et le résultat de cette omission est que les mots tendent à avoir un haut niveau d'ambiguïté. Cette ambiguïté constitue un problème crucial dans la recherche d'information, en fait un mot arabe peut avoir plusieurs significations.

Par exemple, le mot «على» sans voyelles peut signifier le nom propre (Ali) ou la préposition (à).

Dans l'exemple de la figure 2.2, 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.

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

العنوان: اثر العلم. 1- العلماء... 2- علميا... 3- بين العلم الوطني و العلم الأجنبي..	Titre : impact de la science. 1- Les scientifiques…. 2- Scientifiquement… 3- Entre le drapeau national et les drapeaux étrangers …

Fig.2.2: Exemple sur l'effet du mot non voyellé «العلم» sur les extraits.

En plus du phénomène d'ambiguïté, il y a un autre problème de la forme plurielle des noms irréguliers, également appelée le pluriel cassé. Dans ce cas-ci, un nom en pluriel prend une autre forme morphologique différente de sa forme initiale dans le singulier. Par exemple, le mot «امرأة» en singulier prend la forme «نسوة» en pluriel. L'absence d'un dictionnaire pour ces noms irréguliers rend difficile de concevoir un algorithme basé sur les règles pour transformer ce genre de pluriel à la forme singulière.

De plus, les mots sont séparés par des espaces et d'autres signes de ponctuation. Néanmoins, des prépositions sont agglutinées au mot (apparaissant après eux), faisant des limites invisibles entre le mot et la préposition. Par exemple, «أعلمكم».

7. Conclusion

Dans ce chapitre, nous avons présenté certaines caractéristiques de la langue arabe, notamment celles d'ordre morphologique. La langue arabe possède ses propres caractéristiques qui sont différentes par rapport aux langues indo-européennes.

Contrairement aux autres langues, la langue arabe possède un système dérivationnel très riche, et c'est dans cette caractéristique que réside la difficulté de son traitement.

L'Arabe est également très différente des langues européennes au niveau syntaxique. Mais ceci dépasse le cadre de notre travail de recherche parce que l'aspect syntaxique n'est généralement pas pris en compte dans l'état de l'art en recherche d'information. Ainsi, nous nous limitons au traitement morphologique dans ce mémoire.

La langue arabe se caractérise par sa directionnalité droite à gauche, par sa nature cursive (agglutination des mots), par ses signes de vocalisation qui s'ajoutent au-dessous et au-dessus des caractères, et par son ambiguïté due à l'absence de voyelles (cas de la majorité des textes arabes). Ces caractéristiques constituent en fait les problèmes majeurs face aux travaux effectués sur la langue arabe dans le domaine de la recherche d'information.

Pour minimiser ces problèmes nous allons définir quelques traitements sur le corpus de notre étude afin de tenir compte de toutes les spécificités de la langue arabe.

CHAPITRE 3 : PRETRAITEMENTS NECESSAIRES

1. Introduction:

Afin de tenir compte de toutes les spécificités de la langue arabe et de pallier au problème de variation de représentation des caractères arabes dans les textes comme dans les requêtes, il est nécessaire de définir et d'appliquer quelques prétraitements sur le corpus de texte avant l'indexation.

2. Encodage

La langue arabe est encodée suivant plusieurs formats d'encodage comme Unicode, ISO-8859-6, ou autres. Les textes recherchés et les requêtes peuvent être encodés différemment, afin de rendre ceux-ci incomparables. Par exemple, les documents sont représentés en Unicode (UTF-8) et les requêtes en ISO-8859-6 ou un autre encodage. Afin d'apparier les documents avec les requêtes, nous devons réutiliser des outils de conversion entre différents encodages. Ainsi, tout a été transformé en format Unicode dans notre cas.

2.1. L'Unicode:

Le standard Unicode est un mécanisme universel de codage de caractères. Il définit une manière cohérente de coder des textes multilingues et facilite l'échange de données textuelles. Il est obligatoire pour la plupart des protocoles de l'Internet, et mis en œuvre dans tous les systèmes d'exploitation et langages informatiques modernes. Unicode est la base de tout logiciel voulant fonctionner aux quatre coins du monde.

À l'heure actuelle, les données Unicode peuvent être codées sous trois formes principales : une forme codée sur 32 bits (UTF-32), une forme sur 16 bits (UTF-16) et une forme de 8 bits (UTF-8) conçue pour faciliter son utilisation sur les systèmes ASCII préexistants.

2.2. UTF-8

Afin de satisfaire les besoins des systèmes architecturés autour de l'ASCII ou d'autres jeux de caractères à un octet, le standard Unicode définit une forme en mémoire supplémentaire : l'UTF-8.

C'est une forme de mémorisation très fréquemment adoptée pour effectuer la transition des systèmes existants vers Unicode, et elle a notamment été choisie comme forme préférée pour l'internationalisation des protocoles d'Internet.

UTF-8 est un codage constitué de suites d'octets; les bits de poids le plus fort d'un octet indiquent la position de celui-ci dans la suite d'octets.

Autres caractéristiques importantes de l'UTF-8 :

- Conversion efficace à partir de ou vers un texte codé en UTF-16 ou en UTF-32.

- Le premier octet indique le nombre d'octets, ceci permet une analyse rapide du texte vers l'avant.
- Recherche rapide du début de tout caractère, quel que soit l'octet où l'on commence la recherche dans un flux de données, il suffit de consulter au plus quatre octets en amont pour reconnaître aisément le premier octet qui code le caractère.
- UTF-8 est un mécanisme de stockage relativement compact en termes d'octets.

2.3. Produits Unicode supportant l'écriture arabe

Comme nous l'avons vu, l'écriture arabe ainsi que les écritures qui sont dérivées de l'arabe de base se caractérisent notamment par leur directionnalité de droite à gauche, par leur nature cursive et par leurs signes de vocalisation qui s'ajoutent au-dessous et au-dessus des caractères. Ces trois caractéristiques constituent en fait les problèmes majeurs que rencontrent les technologies informatiques. Pour remédier à ces problèmes, le standard Unicode offre toute une panoplie de codes de formatage et des algorithmes permettant, par conséquent, un traitement informatique fiable de l'écriture arabe et des écritures qui en dérivent.

Tirant profit de son importance économique et technologique, le standard Unicode a été vite adopté par un grand nombre de développeurs d'applications et de constructeurs de matériel informatique. Ils exigent que tout nouveau protocole doit être en mesure d'utiliser le codage UTF-8 et que les protocoles existants qui utilisent d'autres jeux de caractères ou même qui utilisent un jeu de caractères par défaut que l'UTF-8, doivent supporter le codage UTF-8.

Le codage Unicode est actuellement supporté par plusieurs systèmes d'exploitation, ainsi que par plusieurs langages de programmation, et par des logiciels.

2.4. L'encodage de corpus et requêtes:

L'arabe est encodé suivant plusieurs formats d'encodage comme Unicode, ISO-8859-6, CP1256, etc. Les textes et les requêtes indexés peuvent être encodés différemment; les rendant incomparables.

Par exemple, si les documents de notre corpus de test sont représentés en ISO-8859-6 et les requêtes, en Unicode (UTF-8), ou en un autre encodage comme Windows CP1256 qui est utilisé sur le Web pour représenter les textes arabes; alors nous devons réutiliser des outils de conversion entre différents encodages en utilisant des tables de l'alphabet arabe, afin d'apparier les documents avec les requêtes. Dans notre cas, tout a été transformé en format Unicode.

3. Segmentation:

Normalement la première étape dans un processus de traitement d'un gros corpus au moyen d'un outil statistique est de subdiviser le texte à traiter en plusieurs unités d'information appelées segments (*tokens)* qui sont, traditionnellement, des mots simples.

3.1. Définition

La segmentation est une étape nécessaire et signifiante dans le traitement de la langue naturelle. La fonction d'un segmenteur est de couper un texte courant en segments, de sorte qu'ils puissent être introduits dans un capteur morphologique ou dans un étiqueteur de position pour un traitement ultérieur. Le segmenteur est responsable de définir des limites de mots, les clitiques délimitantes, les expressions pluri termes, les abréviations et les nombres.

La segmentation est un sujet important dans le traitement de la langue naturelle car elle est étroitement liée « à l'analyse morphologique » [Cha96]. C'est beaucoup plus avec le cas des langues riches et complexes morphologiquement telle que l'arabe. Dans le cas de celle-ci où un mot simple peut comporter un lemme et jusqu'à trois clitiques, la connaissance morphologique doit être incorporée au segmenteur.

3.2. Le système d'écriture arabe:

Le processus de *segmentation* pose une question primordiale : sur le plan informatique, comment repérer un mot ? En d'autres termes, quels sont les indicateurs formels de surface, non ambigus, qui peuvent délimiter un mot ?

En TAL (Traitement Automatique de Langues), on présente généralement les langues, quant à leur système d'écriture, comme appartenant à deux familles différentes : les langues «avec séparateurs» et les langues «sans séparateurs». Les langues dites « avec séparateurs » sont celles qui ont des systèmes d'écritures segmentées c'est-à-dire des écritures délimitées par des espaces, et où les mots sont nettement séparés par des délimiteurs (espace, signes de ponctuation, caractères spéciaux, ...). C'est le cas pour le français ou l'anglais littéraire, ou encore des langues apparentées où la réponse à la question au-dessus est presque trivialle. A ce type de langues on oppose les langues dites « sans séparateurs ». Elles présentent des systèmes d'écritures non segmentées où les mots ne sont pas séparés par des espaces et où les frontières des mots ne sont pas nettes. Dans le cas de termes composés en langue allemande comme, par exemple, *lebensversicherungsgesellschaftsangestellter* ("employé d'une compagnie d'assurance vie"), le japonais, le chinois et le thaï sont aussi les représentants parfaits de cette deuxième famille de langues.

La langue arabe présente un système d'écriture à l'intersection des deux familles. C'est un système d'écriture qui combine une écriture segmentée, et une écriture non segmentée dans laquelle les pronoms sujets et compléments sont dans certains cas attachés aux verbes et une seule chaîne de caractères représente, ainsi une phrase comme par exemple, « كَتَبْتُهُ » ("je l'ai écrit"), cette notion de segments devient carrément inadéquate [Man99].

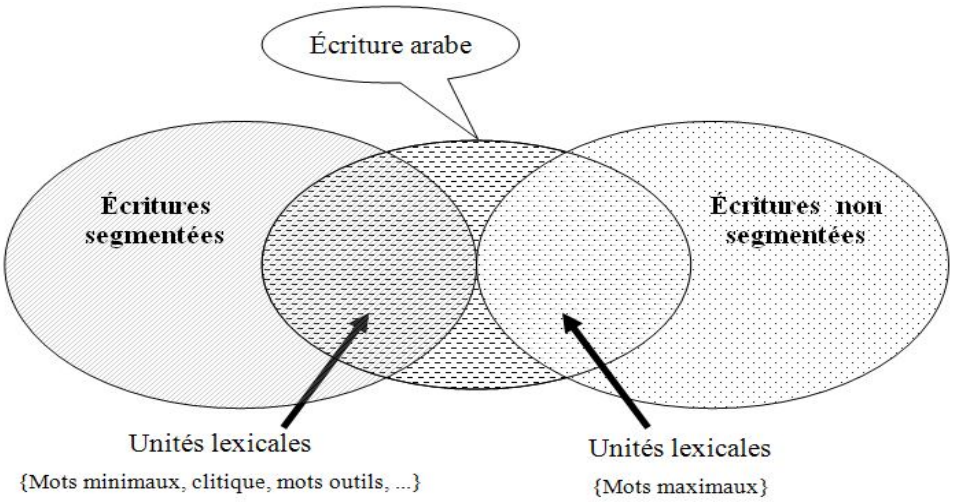

Fig.3.1: Écritures segmentées et non segmentées

3.4. Les types de segmentation

Il existe plusieurs niveaux d'analyse signes d'intérêt permettent de repérer les différents éléments constituant le texte et en définir les frontières. On peut s'arrêter au niveau mot graphique, au niveau des unités lexicales ou aller au delà de celles-ci pour arriver aux unités de base (*les morphèmes*).

Selon la visée de l'analyse à entreprendre : lexicale, morphologique ou syntaxique, on peut généralement trouver trois types de segmentation :

- **La segmentation lexicale** (*tokenization*) qui est la segmentation d'un texte en segments lexicaux (*tokens*). Ce type de segmentation est aussi appelé *itémisation*.
- **La segmentation morphologique** en cherchant à isoler les différents constituants des items lexicaux en unités distinctes, plus petites, qui sont les morphèmes.
- **La segmentation syntaxique** qui consiste à isoler les différents constituants du texte en unités indépendantes, supérieures aux mots, comme les propositions, les syntagmes … etc. Ce type de segmentation est aussi appelé *chunking*.

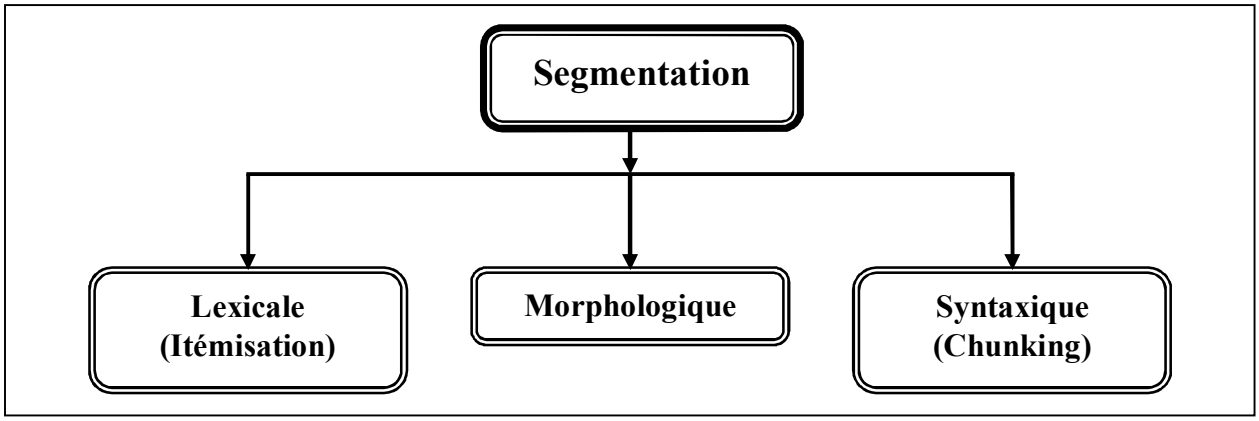

Fig.3.2: Les types de segmentation.

Parmi ces segmentations, nous étendons ici la segmentation lexicale ou itémisation, qui consiste à segmenter un texte en segments ou items lexicaux. C'est une opération consistant à structurer le texte en passant d'un ensemble continu de caractères à une suite discrète d'items lexicaux.

3.5. Les clitiques

Les *clitiques* sont des unités syntaxiques qui n'ont pas des formes libres, mais sont attachés à d'autres mots. La décision si un morphème est un affixe ou un clitique peut être embrouillante. Cependant, nous pouvons généralement dire que les affixes portent les dispositifs morphosyntaxiques (tels que le temps, la personne, le genre ou le nombre), tandis que les clitiques servent les fonctions syntaxique (telles que l'inversion, la définition, la conjonction ou la préposition) qui seraient autrement servies par un élément lexical indépendant. Par conséquent la segmentation est une étape cruciale pour un programme d'analyse syntaxique qui doit construire un arbre à partir des unités syntactiques. Un exemple d'un clitique est la forme contractée "عَلَمْتُهَا"

Les clitiques arabes, cependant, ne sont pas reconnaissables facilement. Ils utilisent le même alphabet que celui des mots, sans la marque délimitante, et ils peuvent être enchaînés l'un après l'autre. Sans connaissance morphologique suffisante, il est impossible de détecter et marquer les clitiques. Ici nous affichons différents niveaux de la mise en place du segmenteur arabe, selon les niveaux de la profondeur linguistique impliqués.

La segmentation arabe a été décrite dans diverses recherches et a été mise en application dans beaucoup de solutions. C'est une étape préliminaire requise pour un traitement plus ultérieur. Ces solutions incluent l'analyse morphologique [Bes01], le diacritique [Nel05], la recherche d'information [Lar02b], et l'étiquetage de position [Dia04] [Hab05]. Aucun de ces projets, cependant, ne décrit la segmentation comme solution autonome ou montré comment l'ambiguïté est filtrée et comment les expressions pluri-terme sont traitées.

3.6. Segments arabes

Un segment est l'unité syntaxique minimale ; il peut être un mot, une partie d'un mot (ou d'un clitique), une expression pluri-terme, ou un signe de ponctuation. Un segmenteur doit connaître une liste de toutes les limites de mots, tels que les espaces et les signes de ponctuation blancs, et également les informations sur les bornes de segment à l'intérieur des mots quand un mot se compose de lemme et des clitiques. Durant la recherche dans la forme complète des mots, c'est-à-dire les lemmes avec ou sans les clitiques, en plus des nombres se nommeront des *segments principaux*. Tous les segments principaux sont délimités par un espace blanc ou un signe de ponctuation. La forme complète des mots peut aussi être divisée en *segments secondaires* où les clitiques et les lemmes sont séparés.

3.6.1. Segments principaux

Un segmenteur se fonde principalement sur les espaces blancs et les signes de ponctuation comme des séparateurs entre les mots (ou des *segments principaux*). Des signes de ponctuation supplémentaires sont utilisés dans l'Arabe tel que la virgule '،', le point d'interrogation '؟' et le point-virgule '؛'. Des nombres sont également considérés en tant que segments principaux. Quelques pays arabes utilisent les chiffres arabes comme en français, alors que la plupart des pays arabes utilisent les chiffres Hindi tels que ٢ '2' et ٥ '5'. Par conséquent une liste de tous les signes de ponctuation et caractères de nombre doit alimenter le système pour lui permettre de délimiter les segments principaux dans le texte.

Entrée	كل وعاء يضيق بما جعل فيه إلا وعاء العلم؛ فإنه يتسع به.
Sortie	كل | وعاء | يضيق | بما | جعل | فيه | إلا | وعاء | العلم | فإنه | يتسع | به

Fig.3.3: exemple sur les segments principaux.

3.6.2. Segments secondaires

Le morpho-tactique arabe permet à des mots d'être préfixés ou suffixés avec les clitiques [Att06]. Les clitiques eux-mêmes peuvent être enchaînées l'un après l'autre, ce qui les rend plus difficiles à manipuler avec n'importe quelle méthode. Un verbe peut comporter jusqu'à quatre segments secondaires (une conjonction, un complément, un lemme de verbe et un pronom d'objet) comme montré par Figure 3.4.

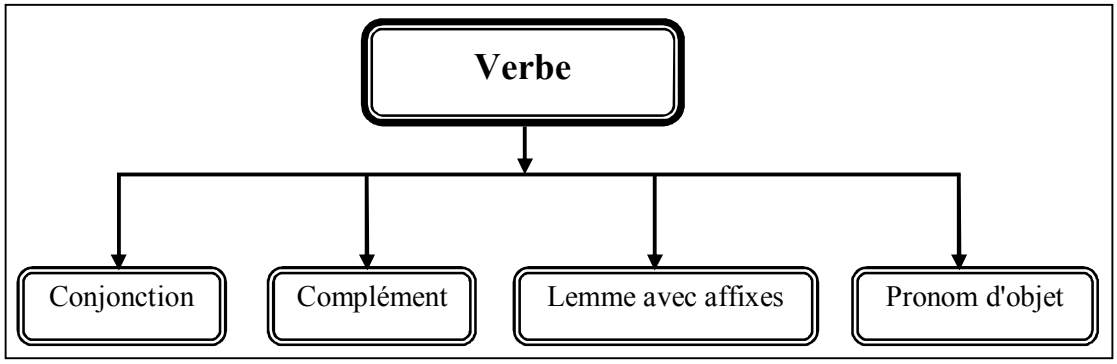

Fig.3.4: Les segments secondaires possibles dans les verbes arabes.

De même un nom peut comporter jusqu'à cinq segments secondaires. Bien que la figure 3.5 montre cinq segments secondaires, l'article défini et le pronom génitif sont mutuellement exclusifs.

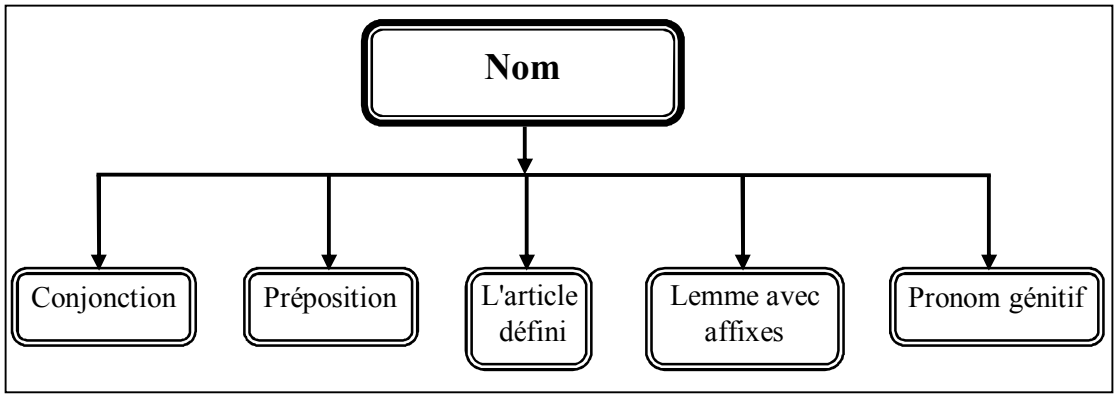

Fig.3.5: Les segments secondaires possibles dans les noms arabes.

De plus, il existe diverses règles qui régissent la combinaison des mots avec des affixes et des clitiques. Ces règles s'appellent les caractéristiques de lexis grammaire [Dic01] [Abb04]. Un exemple sur ces derniers en est la règle qui déclare que des adjectifs et des noms propres ne se combinent pas avec des pronoms possessifs.

3.7. Les solutions de segmentation

Il existe différents niveaux à travers lesquels un segmenteur arabe peut être développé, selon la profondeur de l'analyse linguistique impliquée. Pendant notre travail avec la grammaire arabe nous avons trouvé plusieurs solutions (ou modèles) proposées pour la segmentation arabe. Ces modèles varient considérablement dans leur robustesse, conformité au concept de la modularité, et la capacité d'éviter des ambiguïtés inutiles.

Le segmenteur se fonde sur les espaces blancs et les signes de ponctuation pour délimiter les segments principaux. Cependant, le segmenteur a besoin d'informations plus morphologiques, en délimitant des segments secondaires. Ces informations sont fournies de manière déterministe par un capteur morphologique, ou non déterministe par un guesser (pronostiqueur) de segment. La classification dans les segments principaux et secondaire est une idée conceptuelle qui aide en assignant la tâche de l'identification à différents composants.

L'identification des segments principaux est considérée comme un processus simple qui recherche les espaces et les signes de ponctuation blancs, ensuite segmente le texte suivant ces espaces et ces signes. Aucun autre détail des segments principaux n'est fourni au delà de ce point. Le modèle décrit ci-dessous est la méthode que nous avons choisie pour identifier et diviser des segments secondaire (ou clitiques et lemmes) dans une forme complète de mot.

3.7.1. Le modèle de segmentation: Guesser [Bes03]

Dans ce modèle, la segmentation est séparée de l'analyse morphologique. Le segmenteur détecte alors et délimite seulement des bornes clitiques; pourtant l'information sur ce qui peut constituer un clitique est nécessaire toujours. C'est pourquoi deux composants supplémentaires sont exigés : un guesser (pronostiqueur) de clitiques à intégrer au segmenteur, et un capteur de clitiques à intégrer au capteur morphologique.

3.7.1.1. Le Guesser (pronostiqueur) de clitiques.

Nous avons développé un guesser pour des mots arabes avec tous les clitiques et toutes les assimilations possibles. L'idée de noyau d'un guesser est de supposer qu'un lemme se compose de n'importe quel ordre arbitraire des alphabets arabes, et ce lemme peut être préfixé ou/et suffixé avec un ensemble limité des clitiques. Ce guesser est alors utilisé par le segmenteur pour marquer les bornes de lemme. Les ambiguïtés de segmentation vont s'accroître en nombre à cause de la nature non-déterministe d'un guesser, par exemple,

وللرجل	et à l'homme
#وللرجل#	
#و#للرجل#	
#و#ل#الرجل#	et # à # l'homme #
#و#ل#ال#رجل#	e# à # l # homme #

Fig.3.6: un exemple sur le Guesseur de clitiques.

3.7.1.2. Capteur de Clitiques

Nous devons noter que les clitiques arabes ne se présentent pas séparément dans les textes naturels. Ils sont toujours attachés aux mots. Par conséquent un capteur morphologique de petite taille spécialisé est nécessaire pour manipuler ces formes nouvellement séparées. Nous avons développé un capteur pour les clitiques seulement, pour les retirer ultérieurement facilement. Le but de ce capteur est de fournir l'analyse pour les morphèmes qui ne se présentent pas indépendamment.

Ce capteur spécialisé de petite taille est intégré avec le capteur morphologique principal.

À notre avis c'est le meilleur modèle, et ses avantages sont énormes à un point qu'il peut traiter tous les mots, qu'ils soient connus par le capteur morphologique ou non, et respecter le concept de la modularité car il sépare le processus de segmentation de l'analyse morphologique.

Cependant, ce modèle possède des inconvénients, par exemple les analyseurs morphologique et syntaxique qui doivent traiter l'augmentation des ambiguïtés de segmentation. Le segmenteur est fortement non déterministe car il dépend d'un guesser qui est par définition non déterministe.

4. Les mots vides

Parfois, quelques mots extrêmement communs qui sembleraient être de peu de valeur, peuvent aider à sélectionner les documents correspondant à un besoin de l'utilisateur, sont entièrement exclus du vocabulaire.

Ces mots s'appellent les mots vides. La stratégie générale pour déterminer une liste des mots vides est de trier les termes par fréquence de collection (le nombre total d'apparition de chaque terme dans la collection de documents), et puis de prendre les termes les plus fréquents, souvent filtrés manuellement suivant leur contenu sémantique et en relation avec le domaine des documents indexés, et si nous considérons une liste comme des mots vides alors ses termes sont ensuite rejetés lors de l'indexation. Un exemple d'une liste des mots vides est présenté sur la figure 3.7.

إذ	إذا	اللاتي	اللائي	الذين	الذي	التي	اللتان	اللتين
اللواتي	اللذان	أن	أنا	أنت	أنتما	انتم	انتن	إلى
إليه	إليها	إليهما	إليهم	إلا	أف	الخ	إذن	...

Fig.3.7: un exemple d'une liste des mots vides.

L'utilisation d'une liste des mots vides réduit de manière significative le nombre des termes qu'un système d'indexation doit enregistrer.

Définition: les mots vides sont des mots fréquents qui ne sont pas consultables. Certains moteurs de recherche comprennent des mots tels que «على», tandis que d'autres ignorent ces mots dans une requête.

Dans la plupart du temps, la non-indexation des mots vides ne pose pas un grand problème : les recherches par mots clés avec des termes comme "من" et "إلى" ne semblent pas très utiles. Cependant, cela ne vaut pas pour des recherches d'expression. La requête « الرحلة المتوجهة إلى الجزائر » (vols vers Alger) qui contient un mot vide "إلى" est susceptible d'être perdue si ce mot vide est rejeté. Il y a aussi de nombreux contre-exemples qui concernent le cas de mots vides mal gérés après le déménagement, par exemple:

- *كُنْ أو لا تَكُن*
- *على* الطريق*أيضاً*
- *كل أولئك كانوا في* المدرسة

Les mots en italique sont des mots vides.

Le réglage de la liste des mots vides peut considérablement améliorer les résultats [Sch97]. Créer des listes des mots vides n'est pas généralement considéré comme le traitement automatique des langues naturelles (NLP), mais des techniques du NLP peuvent aider à créer des listes des mots vides et de les traiter.

Nous avons utilisé une liste de mots vides arabes composée des :

- Adverbes,
- Pronoms conditionnels,
- Pronoms interrogatifs.

- Prépositions,
- Pronoms relatifs,
- Pronoms.
- Pronoms verbaux.

Et Autres [Kad08].

Le choix d'un mot quelconque d'entre ces catégories a été basé sur un jugement personnel. Les mots de ces catégories ne sont pas utilisés tous car certains d'entre eux n'ont pas été considérés comme des mots vides.

Ainsi, nous avons enrichi la liste avec d'autres mots vides. Nous avons rassemblé tous ces mots vides dans un fichier dictionnaire nommé MotsVides.dic. Le nombre de mots dans cette collection est égal à 450 mots. Alors que les mots vides de la langue anglaise ou française ne dépassent pas généralement 400 mots. Il y a deux raisons pour lesquelles la liste des mots vides arabe est beaucoup plus grande que les autres listes; d'un coté, les pronoms peuvent avoir plus d'une forme; par exemple, le mot arabe pour ces quatre formes: هاتان (féminin, nominatif), هاتين (féminin, génitif / accusatif), هذان (masculin, nominatif) et هذين (masculin, génitif / accusatif); d'un autre coté, les pronoms et les prépositions sont parfois réunis.

La tendance générale dans les systèmes de recherche d'information avec le temps est l'utilisation standard des grandes listes des mots vides (200-300 termes). Cependant, les moteurs de recherches Web n'utilisent pas des listes des mots vides, alors qu'une partie de la conception des systèmes modernes de recherche d'information s'est concentrée avec précision sur la façon dont nous pouvons exploité les statistiques de la langue afin de mieux pouvoir faire face aux mots communs.

5. Normalisation

Afin de manipuler les variations du texte qui peuvent être représentées en arabe, nous avons appliqué plusieurs genres de normalisation sur le texte et sur les requêtes.

Par exemple, dans l'arabe écrit, les voyelles sont souvent omises dans les textes, néanmoins, on peut parfois trouver quelques voyelles présentes avec les mots. Alors, l'élimination de ces voyelles est nécessaire pour fin de normalisation.

Certaines lettres subissent une simple modification dans l'écriture qui n'influe pas considérablement sur le sens du mot. Mais l'encodage de ces lettres change d'un mot à un autre.

Une autre raison pour ce prétraitement est que l'on a tendance fréquemment à mal écrire ces différentes formes de hamza. Ce genre d'erreurs est très répandu dans les textes arabes. Par exemple, le mot « أكل » est généralement écrit «اكل ». Aussi la lettre « ة » à la fin des mots qui peut être écrite de deux façons : « ة » ou « ه ». Les deux mots arabes « عادة » et « عاده » signifient le même mot (habitude) malgré que leur dernière lettre soit représentée différemment.

La forme normale du corpus a été utilisée pour l'indexation (en conditions non-lemmatisées), et les requêtes ont été normalisées avant qu'elles d'être soumises au moteur de recherche. Des dictionnaires (les fichiers qui contiennent les préfixes, les suffixes, les racines, etc.) ont été également normalisés de sorte que leurs sorties apparient les formes trouvées dans les requêtes et le corpus.

Dans notre démarche, la normalisation a concerné les étapes suivantes :

- Convertir en codage arabe de Windows (CP1256), au besoin
- Enlever la ponctuation
- Retirer les signes diacritiques (principalement voyelles faibles).
- Retirer les non-lettres
- Remplacer le إ ou le أ initial par l'alif nu ا
- Remplacer le آ par le ا
- Remplacer le ىء d'ordre par le ئ
- Remplacer le ى final par le ي
- Remplacer le ة final par le ه

Les définitions de la ponctuation, des signes diacritiques, et des non-lettres sont venues du lemmatiseur de Khoja. Nous allons plus tard améliorer sensiblement la normalisation par l'intermédiaire de deux modifications mineures remplaçant le إ ou le أ par le ا nu d'alif indépendamment de la position dans le mot, et enlever le tatweel.

6. Conclusion

Dans la recherche d'information arabe, le prétraitement ou la préparation de corpus joue un rôle important. Dans ce chapitre, nous avons présenté les prétraitements principaux (l'encodage, la segmentation, la suppression des mots vides et la normalisation) pour chaque corpus de texte arabe.

La langue arabe est encodée suivant plusieurs formats d'encodage comme Unicode, ISO-8859-6, etc. Les textes recherchés et les requêtes peuvent être encodés différemment, les rendant incomparables. Par exemple, les documents sont représentés en Unicode (UTF-8) et les requêtes en ISO-8859-6 ou un autre encodage. Afin d'apparier les documents avec les requêtes, nous devons réutiliser des outils de conversion entre différents encodages. Ainsi, tout a été transformé en format Unicode dans notre cas.

La segmentation est un sujet important dans le traitement de la langue naturelle car elle est étroitement liée à l'analyse morphologique; beaucoup plus avec des langues riches et complexes

morphologiquement telle que l'arabe. Dans le cas de celle-ci, un mot simple peut comporter un lemme et jusqu'à trois clitiques, la connaissance morphologique doit être incorporée au segmenteur.

Les mots vides sont des mots fréquents qui ne sont pas consultables. Ces mots vides peuvent inclure les nombres. Certains moteurs de recherche comprennent des mots vides, tandis que d'autres ignorent ces mots dans une requête.

Afin de manipuler les variations du texte qui peuvent être représentées en arabe, nous avons appliqué plusieurs genres de normalisation sur le texte dans le corpus et dans les requêtes.

Dans le prochain chapitre nous discuterons le cœur de l'analyse morphologique, qui est la lemmatisation, et nous présentons quelques méthodes de lemmatisation.

CHAPITRE 4 : LEMMATISATION

1. Introduction:

L'Arabe est une langue fortement flexionnelle et a une structure morphologique complexe. La recherche d'information sur le texte arabe exige la forme de base du mot (racine ou lemme) pour être la plus pertinente, donc le processus de lemmatisation est nécessaire. La lemmatisation peut être définie comme processus qui consiste à retirer tous les affixes (préfixes, infixes, ou/et suffixes) des mots pour ramener ces derniers à leurs lemmes ou racines. Une racine peut être définie comme le mot qui ne peut pas être créé à partir d'un autre mot, en d'autres termes, un mot sans préfixes, infixes, ou suffixes. Par exemple, la racine du mot arabe (الأعراب, les bédouins) est (عرب, les arabes). Tandis qu'un lemme est défini simplement comme un mot sans préfixe ou/et suffixe. Par exemple, le lemme du mot arabe (الأعرابي, le bédouin) est (اعراب, bédouins). La lemmatisation peut être classifiée, selon le niveau de l'analyse désiré, par exemple : la lemmatisation sur la base du lemme (stem-based) ou sur la base de la racine (root-based).

La recherche d'information, devient de plus en plus sophistiquée; elle tente de répondre à n'importe quelle requête de l'utilisateur. Une de ces tentatives est la lemmatisation de mot. Le problème de disparité du vocabulaire représente un grand obstacle de la recherche d'information sur le texte arabe.

Ce problème se pose quand la forme d'un mot dans une requête n'apparie pas les formes trouvées dans les documents pertinents. Par exemple:

La requête : العرب.

Les documents : ...والعرب، العربي، كالعرب، وعربي

Plusieurs lettres peuvent être attachées à un mot arabe, tandis qu'en français (ou anglais) elles apparaissent en tant que formes séparables. Alors, une requête d'un utilisateur qui contient le mot arabe (العرب, les arabes) n'appariera aucun document qui contienne les mots arabes suivants : (والعرب, et les arabes), (كالعرب, comme les arabes), ...etc.

Il est clair que l'inflexion élevée de la langue aura comme conséquence l'apparition de problème de la disparité du vocabulaire, qui à la suite, réduit significativement l'exactitude de la recherche d'information.

Le processus de recherche d'information est amélioré considérablement quand la lemmatisation est utilisée pour résoudre le problème de disparité de vocabulaire. Pour illustrer l'importance de la lemmatisation dans la recherche d'information, le même exemple précédent est utilisé.

La requête : العرب.

Les documents : ...*عرب، عرب*، **عرب*، ***عرب

Quand la lemmatisation est utilisée pour lemmatiser les mots de la requête de l'utilisateur et des documents, le mot de la requête (العرب, les arabes) sera lemmatisé pour engendrer le

lemme (عرب, arabes). Les mots des documents seront également lemmatisés pour engendrer le lemme (عرب, arabes).

Évidemment, la lemmatisation aide à surmonter le problème de disparité du vocabulaire. Tous les mots dans la requête et les documents ont été lemmatisés au même lemme. Par conséquent, un système de recherche d'information rechercherait tous les documents pertinents. On peut clairement voir que le texte non lemmatisé dégrade la précision de recherche puisque l'arabe est une langue fortement flexionnelle. Ainsi, les lemmatiseurs travaillent généralement sur les langues fortement flexionnelles telles que l'arabe.

2. Définition

La lemmatisation est l'un des nombreux outils utilisés dans la recherche d'information servant à résoudre le problème d'inadéquation vocabulaire des mots; les lemmatiseurs égalisent ou combinent certaines formes variables du même mot comme (papier, papiers) et (pli, plis, plié, pliant…). Dans ce travail, nous utilisons le terme lemme pour se référer à n'importe quel processus qui combine les formes relatives ou groupe des formes dans des classes d'équivalence, y compris l'élimination des affixes.

3. Difficultés de la lemmatisation des mots arabes

La complexité morphologique de la langue arabe rend particulièrement difficile de développer des applications pour le traitement en langue naturel (voir chapitre 2, section 6). Dans les langues sémitiques comme l'arabe, la plupart des lemmes de nom, d'adjectif et de verbe sont dérivés de quelques mille racines par l'insertion de nouvelles lettres, par exemple, les mots مكتبة (bibliothèque), كتاب (livre), كُتُبٌ (livres), كَتَبَ (il a écrit), et نَكْتُبُ (nous écrivons), de la racine كتب [Wig98].

Aussi, la langue arabe est fortement productive, dérivationnelle et flexionnelle. Les articles définis, les conjonctions, les particules et d'autres préfixes peuvent être attachés au début d'un mot et un grand nombre de suffixes peuvent être attachés à la fin d'un mot. Un mot-clé donné peut être trouvé sous différentes formes. Les analyses du texte arabe journalistique prouvent qu'il existe une variabilité lexicologique en arabe plus que dans les langues européenne [ref].

Écrite de droite à gauche, la langue arabe peut aussi être écrite avec ou sans les signes diacritique suivant le contexte de l'apparence des caractères, cette manière d'écriture orthographique contribue également à la variabilité qui peut brouiller les systèmes de recherche de l'information. Par exemple, les deux mots « كَتَبَ » et « كتب » (kataba et ktb) semblent à priori identiques, ce qui n'est pas le même cas pour l'ordinateur, la différence n'apparaît pas. L'orthographe avec des signes diacritiques est moins ambiguë et plus phonétique, mais utilisé seulement dans des contextes spécialisés, tels que les livres d'enfants, les dictionnaires et le Quran. Les signes diacritiques ne sont pas généralement inclus dans les textes tels que les textes des journaux. Pour cette raison, une certaine normalisation, comme la suppression des signes diacritiques, est typiquement représentée dans les systèmes de RI. En revanche, beaucoup

d'analyseurs morphologiques essayent d'insérer les voyelles courtes manquantes et d'autres signes diacritiques.

Aussi, la forme plurielle des noms peut brouiller les systèmes RI. Dans ce cas-ci, un nom au pluriel prend une autre forme morphologique différente de sa forme initiale du singulier. Par exemple, le mot « امرأة » (femme) au singulier prend la forme « نسوة » au pluriel.

En résumé, pour la recherche d'information, cette richesse de formes, lexicologique, et orthographique, rend difficile la correspondance entre la forme d'un mot dans une requête et les formes que l'on peut trouver dans les documents relatifs à notre requête.

4. Les Techniques de lemmatisation

Chaque langue naturelle a ses propres caractéristiques et dispositifs. Ainsi, il semble difficile de suivre la même configuration de lemmatisation et d'appliquer les mêmes techniques pour toutes les langues. Une technique de lemmatisation pourrait être pertinente à une langue, alors qu'elle ne peut effectivement l'être pour d'autres langues, et par conséquent elle ne peut être appliquée. Il existe plusieurs techniques utilisées pour la lemmatisation de mot. Celles-ci incluent des techniques de dictionnaires, d'analyse morphologique, de suppression des affixes, de statistiques et de traduction.

4.1. La technique de dictionnaire

La technique de dictionnaire basée principalement sur la construction d'un dictionnaire très grand en volume qui enregistre les mots trouvés en textes naturels avec leurs parties morphologiques correspondantes [Sug04]. Ces parties incluent : lemmes, racines et affixations; chaque mot utilise une seule entrée dans une table de consultation. Ces dictionnaires sont habituellement construits manuellement, en adoptant une technique où les mots peuvent être lemmatisés par une table de consultation.

Les entrées de table sont mentionnées dans un ordre alphabétique. Une table de hachage ou une liste de recherche binaire (ou dichotomique) peut être utilisée pour optimaliser la recherche.

Les premiers travaux sur la lemmatisation arabe utilisaient les dictionnaires construits manuellement. Al-Kharashi et Evens sont basés sur de petites collections de textes pour lesquelles ils sont construits manuellement des dictionnaires des racines et des lemmes pour que chaque mot soit indexé [Alk94]. Cette approche est évidemment impraticable pour les corpus classés pratiques.

Quoique de telles approches soient précises, les problèmes associés y compris la mise à jour du dictionnaire posent un vrai problème.

4.2. Suppression d'affixe

Dans la langue anglaise et dans d'autres langues d'Europe occidentale, la lemmatisation est principalement un processus de la suppression d'affixes [Lov68] [Por80]. Ce genre de lemmatiseurs ne combine pas les formes irrégulières comme (goose, geese) et (swim, swam).

Ces lemmatiseurs sont généralement taillés selon la spécificité de chaque langue. Leur conception exige l'expertise linguistique et une compréhension des besoins de la recherche d'information. Les lemmatiseurs en général ont été développés pour une large gamme de langues comprenant le Malais [Tai00], le latin [Gre96], l'Indonésien [Ber01] [Gar01], le Néerlandais-suédois [Kra96], l'Allemand [Mon01], le Français [Mou01], le Slovène [Pop92] et le Turc [Ekm96]. L'efficacité de la lemmatisation est variée et influencée par plusieurs facteurs. Généralement la lemmatisation ne cause pas une faiblesse à la recherche ; soit elle fait peu de différence, soit elle améliore l'efficacité [Hul96]. La lemmatisation est considérée comme une aide du rappel plus que de la précision [Kra96], c'est-à-dire, elle permet à un moteur de recherche de trouver des documents plus pertinents, mais ne peut pas améliorer sa capacité d'ordonner le document selon sa pertinence. Elle semble avoir un plus grand résultat positif quand les requêtes et les documents sont petits [Kro93], et quand la langue est fortement fléchie [Pir01] [Pop92], suggérant que la lemmatisation devrait améliorer la recherche d'information arabe.

L'approche de suppression des affixes s'appelle généralement la lemmatisation assouplie ou légère « light stemming », quand elle est appliquée à la langue arabe, elle se réfère à un processus de suppression d'un petit ensemble de préfixes et de suffixes, sans essayer de traiter les infixes, ou d'identifier les modèles (AOUZANE) et de trouver les racines. Cette approche conçue par suppression des chaînes de caractères fréquemment trouvées comme préfixes ou suffixes. Darwish a introduit le lemmatiseur léger Al-Stem à TREC 2002 [Dar02]. Chen et Gey [Car01] ont introduit un lemmatiseur léger, mais celui-ci supprime plus de préfixes et de suffixes et il s'est avéré plus pertinent qu'Al-Stem.

Kadri [Kad08] a proposé une approche[1] similaire aux lemmatiseurs « light stemming ». Elle tronque un mot aux deux extrémités. Le choix des affixes de mots à tronquer est fait selon des statistiques de corpus ainsi que leur rôle syntaxique.

Ce genre de lemmatiseur (suppression d'affixe) peut avoir des points faibles :

- La suppression inexacte de certains affixes. Par exemple, les deux dernières lettres du mot (لسان, langue) peuvent être visualisées comme suffixe (ان) qui est un suffixe duel. Dans ce cas, ce suffixe serait retiré pour produire le lemme incorrect (لس).
- La difficulté d'extraction de la forme singulière d'un pluriel cassé. Par exemple, le lemmatiseur enlève l'affixe attaché au mot (الجداول, les tables) pour produire le mot (جداول, tables). Ce lemmatiseur ne termine pas le processus de lemmatisation par la conversion du mot en sa forme singulière (جدول, table).
- La lemmatisation erronée des mots non arabiques (mots empruntés d'autres langues). Par exemple, le mot (فيتامين, vitamine) est lemmatisé d'une manière erronée pour produire le

[1] Dans cette étude, nous avons combiné cette approche de Kadri avec l'approche basée sur l'analyse morphologique.

lemme « تام » qui signifie complet. Nous pouvons noter que la lemmatisation dans ce cas a compliqué le problème en produisant un mot totalement différent.

4.3. Techniques d'analyse morphologique

Les mots arabes démontrent une morphologie complexe, dont on peut dire que la langue arabe utilise la morphologie racine-modèle où un modèle peut être considéré comme un descripteur adhérant aux règles grammaticales. Ces modèles sont appliqués par l'ajout d'affixes (préfixes, infixes, et suffixes) aux racines (qui sont des verbes nus et ont souvent une longueur de trois lettres) pour former leur racines du parent. Les préfixes et les suffixes peuvent être aussi ajoutés aux lemmes arabes pour exprimer des utilisations grammaticales communes telles que les possessives, les pluriels, les formes définies, le genre, … etc. [Alk94]. Par exemple, certaines formes supplémentaires du mot (كتاب, livre) sont affichées dans le tableau 4.1, où beaucoup de caractères sont attachés au mot (كتاب) tandis qu'en français ces ajouts apparaissent en tant que formes séparées.

Mot arabe	Signification Française
الكتاب	Le livre
كالكتاب	Comme le livre
للكتاب	Pour le livre
بالكتاب	Avec le livre
وكتاب	Et livre

Tableau 4.1 : Certain préfixe attaché au mot livre

La technique d'analyse morphologique est basée sur l'idée de la conformation du mot à un modèle (OUAZENE) pour trouver la racine du mot [Sug04] [Bee98] [Kho99]. La racine est extraite après avoir retiré les affixes attachés à un mot donné. La figure 4.1 illustre le processus complet de l'extraction de racine pour le mot (كاتب, Écrivain) selon le modèle correspondant « فاعل ».

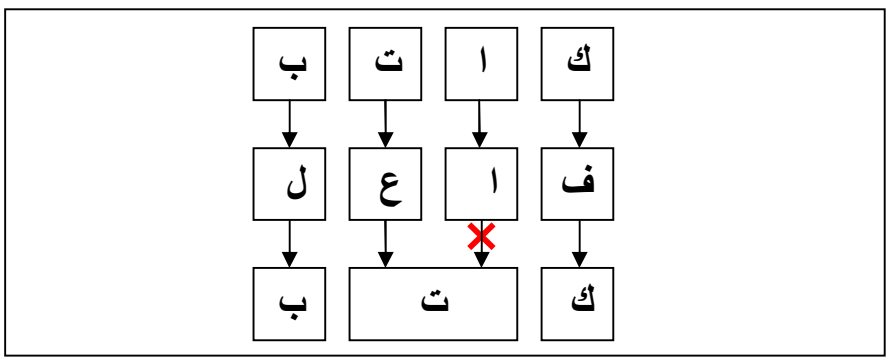

Fig.4.1 : Extraction de la racine du mot « كاتب » du modèle « فاعل »

La majorité des textes arabes standard modernes ont été écrits sans les signes diacritiques (des voyelles ou les signes diacritiques courts), et par conséquent les techniques d'analyse morphologique produisent et posent quelques problèmes tels que :

- La difficulté d'extraire quelques racines. Par exemple, le mot (مخيم, camp) apparie le modèle inadéquat « فعيل » pour produire la racine incorrecte « مخم ».
- La lemmatisation de quelques mots non arabiques d'une manière erronée. Par exemple, le mot (فيروس, virus) apparie le modèle « فيعول », pour produire la racine incorrecte « فرس ».

4.4. Techniques statistiques

Dans cette technique des mesures d'association entre les couples de termes calculées sont basées sur le consécutif de *N* lettres. Les termes qui ont une similitude plus grande (par rapport à un seuil prédéfini) sont groupés et représentés par un terme unique [Sul04] [Xu02]. La mesure de similitude (*S*) basée sur un bigramme unique peut être calculée selon la formule suivante :

$$S = 2 \times C / (A + B) \qquad (4.1)$$

Où :

A : représente le nombre de bigrammes unique dans le premier mot.

B : représente le nombre de bigrammes unique dans le deuxième mot.

C : représente le nombre de bigrammes unique partagés par les deux mots.

Bien que cette technique puisse donner des bons résultats avec certaines langues latines telles que l'anglais, elle poserait quelques problèmes avec les langues fortement inflexionnelles telles que l'Arabe. L'occurrence de ces problèmes est due au fait que la plupart des variantes textuelles de mot impliquent un haut débit de la structure d'infixe, ce qui affecte le calcul des mesures de similitude.

L'exemple suivant montre comment la structure infixe des mots ayant la même signification peut affecter le calcul des mesures de similitude.

Mot1 : (أقوال, paroles) - seuls bigrammes associés : ال، وا، قو، أق

Mot2 : (قول, parole) - seuls bigrammes associés : ول، قو

La mesure de similitude (s) peut être calculée par l'utilisation de l'équation (4.1) comme suit :

$$S= 2\times1/ (4+2)=0.333$$

Bien que les deux mots de l'exemple soient très semblables et appartiennent à la même catégorie, leur coefficient de similitude calculé par cette technique est très bas et loin de la valeur de seuil acceptable.

L'exemple suivant montre la valeur de similitude des mots (sans infixe) ayant une signification différente :

Mot1 : (مكتب, bureau) - seuls bigrammes associés : تب، كت، مك

Mot2 : (مكتبة, bibliothèque) - seuls bigrammes associés : بة، تب، كت، مك

$$S= 2\times3/ (3+4)= 0.8571$$

En outre, la technique statistique de N-gramme ne peut pas être pertinente aux langues qui ont une inflexion élevée parce que la valeur de similitude des mots qui ont une signification différente peut être assez grande pour grouper ces mots comme semblables.

4.5. Techniques de traduction

La traduction est une technique utilisée pour appliquer des approches déjà développées pour des langues latines telles que l'anglais sur des langues fortement flexionnelles comme l'Arabe. L'inflexion élevée de quelques langues peut être réduite par une application de cette technique [Che02].

Cette technique pose quelques problèmes, tels que :

- Les traductions ambiguës, par exemple, la traduction des deux mots arabes différents tels que (مريض, patient) et (صبور, patient) au même mot français « patient », ainsi que la traduction de mot arabe (ذهب, or) à un mot français incorrect « aller ».
- L'efficacité des algorithmes de lemmatisation avec les langues latines elles-mêmes : Certains lemmatiseurs tels que Porter échoue à lemmatiser quelques mots. Par exemple, le mot « children » en anglais (أطفال, enfants) n'est pas lemmatisé par Porter pour engendrer le mot « child » (طفل, enfant).
- L'échec de choisir le lemme correct : Cette technique choisit le plus court terme parmi un ensemble des termes groupés pour représenter un lemme candidat. Dans certains cas, cette technique rencontre une difficulté pour choisir un lemme candidat. Par exemple, les mots (وثيقة, document) et (وثائق, documents) tous les deux ont la même longueur.

5. La méthode proposée

Dans ce travail, nous avons proposé une méthode hybride qui incorpore trois techniques différentes pour que la lemmatisation arabe résolve les problèmes liés à chacune des techniques précédentes.

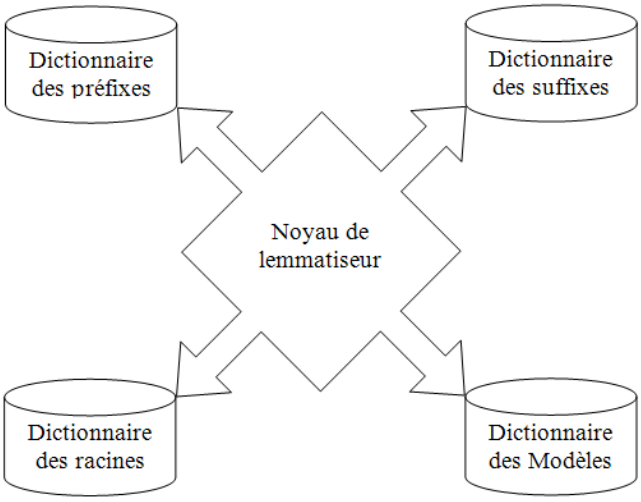

Fig.4.2 : Architecture de Notre lemmatiseur.

Les trois techniques sont: suppression d'affixe proposée par Kadri [Kad08], dictionnaires et analyse morphologique.

Ces techniques ont besoin d'une certaine adaptation pour être pertinentes pour l'utilisation. Chaque technique est adaptée individuellement pour résoudre les problèmes pratiques liés à elle-même. Les sections suivantes décrivent en détails les techniques intégrées dans la méthode proposée.

5.1. Suppression d'affixe

Cette technique commence par la détermination de tous les affixes possibles (préfixes et suffixes) qui peuvent être attachés aux mots arabes.

La liste des préfixes et des suffixes de la langue arabe étant limitée, on peut utiliser celle proposée par [Abb04] pour la lemmatisation.

Préfixe	Fréquence	Préfixe	Fréquence	Préfixe	Fréquence	Préfixe	Fréquence
ا	13324	است	298	ولل	71	ولن	8
و	10232	من	291	مه	64	فست	7
ال	9965	مس	288	وسي	64	فسا	6
وا	4475	مت	277	مم	61	فلي	6
ب	4040	ه	269	فل	59	اا	5
ل	3821	ون	243	الال	52	اسن	5
م	3344	سي	233	لك	52	فبال	4
وال	3315	في	192	ولا	52	مته	4
ت	3167	فت	188	سن	49	متي	4
ي	2040	لي	184	وبال	48	وكال	4
ف	1491	ست	177	وست	47	ولب	4
لل	1417	كال	172	فن	33	افن	3
الا	1323	ام	162	اسا	29	فلن	3
بال	1257	يس	160	فب	29	افسا	2
وت	1233	ين	148	فك	24	افي	2
ك	987	اب	135	ولي	23	فلت	2
ن	930	وك	134	متن	22	اكال	1
الت	816	مست	124	ولت	22	فسن	1
وي	628	سا	122	وسا	21	ابال	0
لا	477	تم	119	متم	17	افست	0
فا	466	أي	118	ولك	17	افسن	0
ان	432	ات	105	افت	15	افسي	0
لت	429	مي	102	اسي	12	فكال	0
وب	360	اك	99	فلا	11	فلال	0
الن	334	الي	92	لال	11	لبال	0
يت	322	لن	84	وسن	10	لكال	0
فال	313	لو	80	افا	8	ولال	0
ول	311	لب	73	فسي	8	ولبال	0

Tableau 4.2 : Fréquences d'occurrence des préfixes sur les mots de la collection «Al-Khat Alakhdar».

L'approche de Kadri est similaire aux lemmatiseurs utilisés généralement sous l'appellation «light stemming». Elle lemmatise un mot aux deux extrémités. Le choix des affixes de mots à lemmatiser est fait selon des statistiques de corpus ainsi que leur rôle syntaxique. Nous avons groupé tous les affixes dans deux classes : préfixes et suffixes, basées sur les fréquences d'occurrence de ces affixes sur les 50172 mots différents de la collection arabe «Al-Khat Alakhdar». Les tableaux 4.2 et 4.3 résument successivement ces statistiques.

Suffixe	Fréquence	Suffixe	Fréquence	Suffixe	Fréquence	Suffixe	Fréquence
ه	10550	وها	40	ناك	3	تكما	0
ا	6900	وك	38	وننا	3	تمانا	0
ت	4660	هن	35	وهن	3	تمانى	0
ن	3898	ونى	35	يكن	3	تماه	0
ى	3297	اتى	34	يننا	3	تماها	0
ها	3199	يني	34	يهن	3	تماهم	0
ات	3128	نك	30	اتهما	2	تماهما	0
يه	2799	الم	29	اتهن	2	تماهن	0
م	2507	ناه	29	تكن	2	تمايا	0
ان	936	ننا	27	ماها	2	تمونا	0
ون	936	اتنا	26	مونا	2	تمونى	0
هم	803	ونا	24	نهن	2	تموه	0
تها	752	ينها	24	اكما	1	تموهم	0
ك	648	تن	19	انهن	1	تموهما	0
ته	621	اكن	16	تانا	1	تموهن	0
نا	598	تهما	16	تاهما	1	تننا	0
و	494	وهم	16	تما	1	تنني	0
نه	441	يهما	16	تموها	1	تنهم	0
وا	436	مانى	15	تنها	1	تنهما	0
نى	367	ناها	14	تيها	1	تنهن	0
اه	286	تيك	13	تيهم	1	تيكم	0
اتها	232	تني	12	موهم	1	تيكما	0
تهم	225	اننا	10	ناهم	1	تيكن	0
يها	225	ماه	9	ناهما	1	تيهما	0
نها	196	نهما	9	نكم	1	تيهن	0
تى	139	اتك	8	نكن	1	كم	0
اته	138	اهما	8	وكم	1	كن	0
وه	137	موه	8	وكما	1	ماهم	0
انى	131	تانى	7	وكن	1	ماهما	0
اك	124	تكم	7	ونك	1	ماهن	0
تان	114	موها	7	ونهما	1	موهما	0
تا	109	ينهم	7	وهما	1	موهن	0
اها	108	كما	6	يكما	1	ناكم	0
يك	107	يكم	6	ينهما	1	ناكما	0
اتهم	85	انك	5	اتكما	0	ناكن	0
تنا	83	ننى	5	اتكن	0	ناهن	0
ينا	75	تاك	4	انكم	0	نكما	0
ينه	75	تهن	4	انكما	0	هما	0
ونه	73	مانا	4	انكن	0	ونكم	0
انا	61	مونى	4	ة	0	ونكما	0
تيه	60	ونهم	4	تاكم	0	ونكن	0
انها	54	اتكم	3	تاكما	0	وننى	0
نهم	54	اننى	3	تاكن	0	ونهن	0
تم	53	انهما	3	تاها	0	ية	0
يهم	52	اهن	3	تاهم	0	ين	0
ونها	50	تاه	3	تاهن	0	يننى	0
اى	48	تينا	3	تاى	0	ينهن	0

Tableau 4.3 : Fréquences d'occurrence des suffixes sur les mots de la collection «Al-Khat Alakhdar».

Finalement, nous avons établi les listes de préfixes et de suffixes les plus fréquents pour lemmatiser des mots. Ce choix est aussi guidé par le rôle syntaxique de ces affixes dans les textes arabes. La liste des préfixes à supprimer que nous avons établie est la suivante :

ا , و, ال, وا, ب, ل, م, وال, لل, الا, بال, فال, ول, است, كال, ولل, الال, فل, وبال, وست, مست.

Les suffixes que nous avons jugés nécessaires de lemmatiser étant les plus fréquents et représentent généralement des pronoms attachés à la fin des mots expriment le nombre ou le genre des noms arabes [Kho99]:

ه , ا, ت, ن, ي, ها, ات, يه, م, ان, ون, هم, تها, ك, ته, نا, و, نه, وا, ني, اه, اتها, تهم, يها, نها, تي, اته, وه, اني, اك,
تان, اها, يك.

En résumé, nous avons classé dans notre travail tous les préfixes et suffixes selon leur rôle syntaxique quand ils sont attachés aux mots arabes.

D'autre part, nous avons donné une argumentation linguistique et statistique pour choisir les préfixes et les suffixes. Nous avons choisi les préfixes qui sont généralement des prépositions attachées aux débuts des mots, les suffixes qui sont des pronoms collés à la fin des mots.

Finalement, et pour la fiabilité, nous avons spécifié une valeur convenable pour la longueur du mot traité. Par exemple, une valeur de 2 est spécifiée pour tous les affixes. En d'autres termes, si un mot est attaché avec des affixes, nous devons contrôler la longueur de ce mot sans ces affixes, c'est-à-dire si la longueur est supérieure ou égale à 2, nous pouvons supprimer les affixes, sinon, aucune élimination ne serait effectuée. Par exemple, le mot (سهم, flèche) pourrait être incorrectement lemmatisé en retirant les deux dernières lettres « هم » pour produire un lemme sans signification « س ». Dans un autre exemple, le mot (ربهم, leur seigneur), selon la méthode proposée en retirant le suffixe (هم, leur) pour produire le mot (رب, seigneur). Dans ce cas-ci, la longueur du mot examiné est contrôlée et si elle satisfaite la condition (supérieur ou égale à 2), la lemmatisation serait appliquée.

5.2. La technique de dictionnaire

Cette technique est adoptée pour atteindre deux objectifs de lemmatisation. Le premier est de résoudre la suppression incorrecte de quelques affixes, tandis que le second est de traiter le problème des mots arabisés (Arabicized).

Le premier but est atteint par la construction d'un dictionnaire de racine contenant les racines des mots Arabes. Ces racines sont librement disponibles sur l'Internet pour les recherches[1]. Pour lemmatiser un mot donné, ce dernier traverse une série d'étapes.

Ces étapes sont récapitulées comme suit :

étape1: Vérifier le mot donné dans le dictionnaire, s'il n'existe pas, on passe à l'étape 2, sinon renvoyer le mot en tant que racine.

étape2: Retirer tous les affixes possibles attachés au mot, ensuite vérifier le mot dans le dictionnaire, s'il n'existe pas, on passe à l'étape 3, sinon on renvoie le mot en tant que racine.

étape3: Le mot résultant obtenu à partir de l'étape 2 est comparé a un ensemble de modèles (paternes) pour extraire sa racine, si la racine du mot résultant n'existe pas dans le dictionnaire ou n'est appariée avec aucun modèle, alors le mot peut être erroné et il faut le sauvegarder tel qu'il est.

---

[1] Dr. Salim al-Khammash «Lisan al-Arab». King Abdulaziz University.
http://www.angelfire.com/tx4/lisan/roots1.htm, accede le 05/12/2008.

ابا	بابا	تاتا	ثاثا	جاجا	حاحا	خبا	دادا
اتا	بتا	تاب	ثدا	جبا	حبا	ختا	دبا
اثا	بثا	تار	ثرطا	جرا	حتا	خجا	دثا
اجا	بدا	تاف	ثطا	جزا	حجا	خذا	درا
اشا	بذا	تاق	ثفا	جسا	حدا	خرا	دفا
الا	برا	تال	ثما	جشا	حزا	خسا	دكا
اوا	بسا	تالب	ثاب	جفا	حشا	خطا	دنا
ابب	بطا	تام	ثبب	جلا	حصا	خفا	دهدا
اتب	بكا	تان	ثرب	جلظا	حضا	خلا	دوا
اثب	بها	تاي	ثرقب	جما	حطا	خما	داب
ادب	بوا	تبا	ثعب	جنا	حبطا	خبب	دبب

Tableau 4.4 : un aperçu sur le dictionnaire des racines (DJOUDHOUR).

Pour atteindre le deuxième objectif, on a construit un dictionnaire des mots arabisés (Arabicized).

Ces mots sont rassemblés d'un livre arabe [Tou05]. La lemmatisation des mots arabisés passeront par les étapes suivantes:

étape1: Vérifier le mot donné dans le dictionnaire des mots arabisés, s'il n'existe pas, passer à l'étape 2, sinon le mot est reconnu en tant qu'un mot arabisé.

étape2: Supprimer (un par un) les affixes possibles du mot, et vérifier à chaque fois le mot résultant dans le dictionnaire, s'il n'existe pas, alors le mot peut être erroné et il faut le sauvegarder tel qu'il est, autrement le mot est reconnu en tant qu'un mot arabisé.

5.3. Techniques d'analyse morphologique

La technique d'analyse morphologique est utilisée pour atteindre deux objectifs de lemmatisation. Le premier objectif est de diminuer le nombre de cas de la suppression incorrecte de quelques affixes (qui ont des lettres principales qui apparaissent en tant qu'affixes). Tandis que le deuxième objectif est d'aider à convertir les mots pluriels cassés en leurs formes singulières.

Le premier objectif peut être atteint en utilisant un ensemble de modèles arabes pour améliorer le processus de lemmatisation.

استفال	أفاعلة	فئعلان	فعئل	فعلئ	فعملة
استفع	أفاعيل	فاع	فعا	فعلئة	فعمولة
استفعال	أفع	فاعال	فعائل	فعلاء	فعنال
استفعل	أفعال	فاعل	فعاة	فعلاءة	فعنالة
استفعى	أفعالة	فاعلاء	فعاعيل	فعلات	فعنعل
استفل	أفعالون	فاعلاع	فعافع	فعلاة	فعنفال
استفيل	أفعالية	فاعلال	فعافعة	فعلاع	فعنفع

Tableau 4.5 : un aperçu sur les modèle (AOUZANE) de notre travail.

Pour traiter le problème de la suppression incorrecte de quelques affixes, le mot concerné est comparé avec un ensemble de modèles. Si un appariement a eu lieu, alors le processus de suppression d'affixe sera annulé parce que l'affixe est considéré en tant qu'élément du mot.

Le deuxième objectif principal de la lemmatisation est la conversion d'un pluriel à son singulier associé. L'application de la lemmatisation sur des pluriels irréguliers (pluriels cassés) est une tâche difficile parce que ces types de pluriels ne suivent pas des règles grammaticales évidentes.

Pour lemmatiser des pluriels cassés, nous devons d'abord identifier correctement les formes des pluriels cassés à partir d'autres formes de mots. L'application des modèles de pluriels cassés seulement ne garantit pas l'identification réussie et efficace. Par exemple, le mot (أوزان, poids) apparie le modèle pluriel cassé (أفعال), et en conséquence, il est correctement identifié comme pluriel cassé.

Inversement, le mot (إجحاف, préjudice) apparie le modèle pluriel cassé (افعال), mais il n'est pas un pluriel cassé. Ce type de problème se produit à cause du manque d'information.

Une comparaison simple des mots avec les modèles pluriels cassés aide à identifier des pluriels cassés avec une exactitude acceptable; cependant, elle inclut un certain bruit par la classification d'autres mots en tant que pluriels cassés.

Parfois un pluriel cassé possède plusieurs formes singulières, ce qui cause une difficulté pour choisir le bon singulier, par exemple les mots (وسائل, moyens), (رسائل, messages) et (زبائن, clients), ces trois mots correspondent au modèle « فعائل » en pluriel cassé, ce dernier a trois modèles formes singulières différentes : « فعيلة », « فعالة », « فعول ». Le problème que nous avons rencontré est que nous ne pouvons pas choisir le modèle singulier adéquat. Il n'y a aucune règle ou mécanisme que nous pourrions suivre pour choisir le bon modèle singulier. Dans notre cas, le premier mot (وسائل, moyens) est converti à son singulier (وسيلة, moyen) qui apparie le modèle singulier « فعيلة », le deuxième mot (رسائل, messages) est converti à son singulier (رسالة, message) qui apparie le modèle singulier « فعالة », et le troisième mot (زبائن, clients) est converti à son singulier (زبون, client) qui apparie le modèle singulier « فعول ».

6. Conclusion

Le traitement morphologique est le cœur de la recherche d'information arabe et plus précisément la lemmatisation qui joue un rôle important. Cette dernière a fait l'objet de plusieurs travaux de recherches.

La lemmatisation est importante pour les langues fortement flexionnelles telles que l'Arabe pour beaucoup d'applications qui exigent le lemme d'un mot. Les lemmatiseurs arabes existants ne produisent pas des résultats parfaitement pertinents. Dans ce travail, nous avons proposé une méthode de lemmatisation hybride qui essaye de déterminer le lemme d'un mot selon des règles linguistiques. La méthode proposée intègre trois techniques différentes pour améliorer la performance globale du processus de lemmatisation.

Notre méthode présente une intéressante performance de recherche que les autres méthodes (voir chapitre 5, section 4), parce qu'elle permet de mieux déterminer le lemme d'un mot, Alors que les autres méthodes ne permettent pas avec réussite de grouper beaucoup de mots sémantiquement similaires dans le même index.

Dans le prochain chapitre nous présentons l'implémentation, l'expérimentation, et l'évaluation de notre système d'indexation et de recherche des textes en arabe, et nous montrons nos contributions dans cette étude.

CHAPITRE 5 : IMPLEMENTATION ET EXPÉRIMENATION

1. Introduction:

La recherche d'information en langue arabe est devenue de plus en plus importante. Ce domaine de recherche actif connaît un grand progrès ces dernières décennies. A cette fin, nous nous sommes intéressés au traitement automatique des langues naturelles.

Nous avons réalisé un système de recherche d'informations dédié à la langue arabe fondé sur une méthode hybride qui combine trois techniques connues : suppression d'affixe, dictionnaires, et analyse morphologique. Ce système a été évalué sur un corpus, que nous avons construit, traitant la thématique d'environnement[1].

Dans ce chapitre nous allons présenter notre travail et contribution dans cette étude. En effet, le but de notre travail est d'étudier les différentes méthodes de recherche d'information en langue arabe, d'appliquer quelques méthodes d'indexation ou de lemmatisation, de comparer les résultats et conclure à propos de la méthode qui donne une meilleure performance dans la recherche d'information.

2. Le corpus de test:

Pour démontrer l'intérêt de représenter le contenu textuel par des unités lexicales dans un processus de recherche d'information, nous devons disposer d'un corpus de langue arabe riche en termes de variation de genres. A notre connaissance, la collection TREC incluant des documents, des requêtes et des jugements de pertinence est la plus grande collection en arabe actuellement disponible. Elle contient 383 872 articles provenant d'Arabic Newswire de l'AFP (Agence France Presse). Ainsi la collection représente un volume de 884 MOctets. Ces articles sont des articles de journaux arabes couvrant la période de mai 1994 jusqu'à décembre 2000 [Kad08], et comme nous ne disposons pas de cette collection ou d'une autre qui soit professionnelle et qui nous aide dans le processus de la recherche. Nous avons décidé de construire un corpus à partir du web. Pour collecter des documents, nous avons effectué une recherche sur le web à l'aide des moteurs de recherche, et avons trouvé la collection « Al-Khat Alakhdar » [1]. Cette dernière, spécialisée dans le domaine de l'environnement, est restreinte aux thématiques suivantes : la pollution, la purification de l'eau, la dégradation du sol, la préservation de la forêt, les catastrophes naturelles…etc. Cette collection, Al-Khat Alakhdar, dont nous avons intégré quelques requêtes et jugements de pertinences fait l'objet d'une importante production langagière en arabe contient 694 articles couvrant la période de septembre 2001 jusqu'à ce jour, représentant un volume de 15 MOctets.

[1] http:// www.greenline.com.kw. Accédé le 14/09/2008.

الخط الأخضر تدعو لإيجاد أنماط بيئية للاستهلاك

دعت جماعة الخط الأخضر البيئية في بيان صادر لها مجلس الأمة والحكومة إلى ضرورة إيجاد رؤية واضحة للتعامل مع أنماط الاستهلاك غير المستدامة الموجودة في الكويت خصوصا في ظل التزايد السكاني المستمر وعدم اتزان التركيبة السكانية . وذكرت جماعة الخط الأخضر البيئية في بيانها الصادر بمناسبة اليوم العالمي للسكان أن عدم تعامل الجهات الحكومية المسئولة بشكل صحيح مع حجم النفايات المتولدة من الممارسات اليومية للسكان في دولة الكويت تنذر بكوارث بيئية وصحية واقتصادية ستؤدي لنسف أية خطط إستراتيجية تنموية .

وحذرت جماعة الخط الأخضر البيئية من الفراغ التشريعي المتمثل بعدم وجود قانون بيئي يحمي سكان الكويت البالغ عددهم ما يقرب من الثلاثة ملايين نسمة من مخاطر التلوث الذي بات يفتك بصحتهم وصحة أطفالهم ، كما لا ترتقي الخدمات الصحية في الدولة إلى المستوى الذي يمكن أن يعالج الأمراض الناتجة عن التلوث الذي تعاني منها البلاد . وأكدت جماعة الخط الأخضر البيئية أن السياسات الاقتصادية الحكومية تفتقد إلى الرؤية الواضحة القائمة على أسس التنمية المستدامة التي تضمن سلامة السكان وضمان الأمن البيئي والصحي لأفراد المجتمع. واستغربت جماعة الخط الأخضر البيئية ندرة الدراسات البيئية التي تبحث أسس التعامل مع الزيادة السكانية والبيئة ، خصوصا في ظل وجود اكثر من مؤسسة بحثية حكومية.

هذا وأكدت جماعة الخط الأخضر البيئية أن استشراء الأمراض الناتجة عن التلوث بين أفراد المجتمع الكويتي أتى كنتيجة رئيسية لسوء التعامل الحكومي مع الزيادة السكانية المستمرة. وأضافت جماعة الخط الأخضر البيئية إلى أن الزيادة المستمرة في عدد السكان تستلزم من الدولة فهم واسع لكيفية التعامل الصحيح مع هذه الزيادة للاستفادة الايجابية منها خصوصا وأن ذلك يعني إنشاء مدن جديدة ومشاريع خدمية كبيرة الأمر الذي يتطلب مراعاة الاشتراطات البيئية أثناء تنفيذ هذه المشاريع الإسكانية والخدمية حتى لا تتكرر الأخطاء التي ارتكبت سابقاً.

وأكدت الجماعة إلى أن الحاجة ماسة جدا إلى إيجاد برنامج شامل وعام للارتقاء بمستوى الوعي البيئي لدى جميع أفراد المجتمع وذلك لتغير أنماط الاستهلاك وسلوك الفرد في التعامل مع بيئته. هذا وحذرت جماعة الخط الأخضر البيئية من أن عدم قيام الهيئة العامة للبيئة بدورها المفترض سيؤدي إلى تفاقم المشاكل والجرائم البيئية وانتهاكات حقوق الإنسان البيئية بحق المجتمع . كما دعت إلى ضرورة تفعيل وتسهيل دور القطاع الخاص في سبيل إنشاء مصانع بيئية متخصصة في تدوير النفايات المختلفة لتقليص حجم المساحات المستخدمة لردم النفايات والاستفادة من المردود الاقتصادي لعمليات تدوير النفايات .

واختتمت الجماعة بيانها بدعوة الحكومة إلى بناء منظور إستراتيجي لتحويل المجتمع من مجتمع استهلاكي غير مستدام إلى مجتمع يجيد التعامل مع المتغيرات المحيطة به بما يخدم بيئته وصحته.

التاريخ: 10 يوليو 2008

Fig.5.1 : Exemple d'un document arabe dans la collection «Al-Khat Alakhdar»

Pour nos expérimentations, un corpus (documents et requêtes) a été construit en s'inspirant des campagnes d'évaluation TREC. Cette forme apporte une information aussi complète et détaillée que possible, y compris des connaissances avancées sur le domaine grâce aux mots-clés. Un exemple de ces documents est présenté sur la figure 5.2.

<DOC>
<DOCNO> AR-017 </DOCNO>
<HEADLINE>
مساحات كبيرة من الغابات تختفي كل عام
<HEADLINE/>
<TEXT>
علماء يتابعون مصير 3 ملايين شجرة في محمية بيولوجية في قناة بنما
بنما:"نيويورك تايمز" في عام 1979 توصل عالما بيئة في جامعات في الغرب الأوسط في الولايات المتحدة، يعرفان بعضهما البعض عبر الأبحاث الى فكرة جريئة. وأرادا الحصول على حقوق شاملة لإجراء الأبحاث من قمة جزيرة بارو كولورادو المخصصة للأبحاث التي أصبحت واحدة من أكثر الأماكن دراسة في العالم. والجزيرة، وهي عبارة عن محمية بيولوجية في قناة بنما، يديرها معهد الأبحاث الاستوائية التابع لمعهد سميثونيان. ولذا قرر العالمان روبين فوستر الذي كان في جامعة شيكاغو انذاك، وستيفن هابل الذي كان في جامعة ايوا، الاتصال بمدير المعهد ايرا ربونوف، واقترحا اجراء مسح شامل وقياس جميع الاشجار في الجزيرة كل خمس سنوات لمعرفة التغييرات واجراء تجارب على النظريات المتعارضة حول تنوع الغابات الاستوائية.
<TEXT/>
<DOC/>

Fig.5.2 : Exemple de document.

Un autre exemple de ces requêtes est présenté dans la figure 5.3.

<REQ>
<REQNO>AR-05</REQNO>
<TITLE> حماية الغابات <TITLE/>
<DESC> النصوص التي تتحدث عن حماية الغابات <DESC/>
<NARR> النصوص التي تتحدث عن رعاية الغابات و الاهتمام بها، والنهي عن التعدي على الغابات عن طريق تعطيش أو تقطيع أشجار الغابة، وغيرها من الطرق الغير مشروعة<NARR/>
<REQ/>

Fig.5.3 : Exemple de requête.

Le tableau 5.1 présente quelques caractéristiques de la collection «Al-Khat Alakhdar».

Langue du corpus des documents	**arabe**
Nombre de documents	**694**
taille du corpus (MB)	**15**
Nombre total de mots (tokens)	**412407**
Nombre de mots différents	**50172**
Taille moyenne des documents (mots)	**595**
Langues des requêtes	**arabe**
Nombre de requêtes	**10**
Taille moyenne des requêtes (mots)	**3**

Tableau 5.1 : Caractéristiques de la collection arabe «Al-Khat Alakhdar»

3. Implémentation:

Parmi nos objectifs est l'utilisation de la langue arabe dans le domaine d'indexation qui serait un pas supplémentaire vers son intégration dans la technologie de l'information vue sa puissance et sa richesse. Nous nous sommes basés dans notre implémentation d'un système RI sur deux grands axes : l'indexation et la recherche.

3.1. Indexation

Soit les notations suivantes :

$\mathcal{C}$: un corpus ou un ensemble de documents $\{d_1, d_2, ... d_N\}$.

N: le nombre de documents du corpus.

$\boldsymbol{d_i}$: un document ou une séquence de termes t, notée $<t_1, t_2, ...t_i>$.

$\boldsymbol{l_i}$: la longueur du document d_i

$\mathcal{T}$: le dictionnaire ou l'ensemble des termes distincts du corpus $\mathcal{C}$.

$\boldsymbol{idx_j} = (o_1, o_2, ...o_N)$: l'index du terme t_j pour le corpus $\mathcal{C}$ où o_k définit le nombre d'occurrences du terme t_j dans le document d_k.

$\boldsymbol{pos_{jk}}=(p_1, p_2, ...p_{ok})$: le vecteur des positions du terme t_j dans le document d_k où p_m définit la position de la $m^{ième}$ occurrence du terme t_j dans le document d_k

Un moteur d'indexation est le calcul de idx_j et de pos_{jk} pour tous les termes t_j du dictionnaire $\mathcal{T}$ du corpus $\mathcal{C}$. [Ben06]

Ou bien, tout simplement, l'indexation est le processus qui permet de représenter un document $\boldsymbol{d_i}$ pour le rendre exploitable d'une manière efficace par une recherche ultérieure.

L'indexation est définie par AFNOR[1] comme un processus destiné à représenter par les éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou d'une question. On désigne également ainsi le résultat de cette opération [AFN93].

3.2. Recherche d'information

La recherche d'information est fortement liée à l'indexation. En effet, à quoi cela sert-il d'indexer des textes si les informations et leurs emplacements repérés ne sont pas réutilisés par un système de recherche ?

L'objectif de la recherche des documents est de ressortir les documents les plus pertinents de la collection pour une bonne interprétation.

---

[1] **AFNOR**: l'Association Française de normalisation.

La réponse à une requête cherchant les documents qui contiennent le terme t_Q est directement obtenue avec idx_Q, l'index du terme t_Q. Si $o_k \neq 0$, cela indique la présence du terme recherché dans le document d_k. Le vecteur des positions pos_{jk} est nécessaire dans le cas où la requête stipule un rapport de distance ou de précédence entre deux termes recherchés.

Le processus de recherche dans le résultat de l'indexation est le suivant :

- Rechercher les identifiants des termes de la requête dans le dictionnaire.
- Rechercher les index des termes.
- Filtrer, ordonnancer le résultat.
- Rechercher les noms des documents du résultat.

3.3. Architecture du système

Notre système permet de fournir au corpus de test, l'index possible et permet la recherche tout en optimisant les coûts en termes de temps et d'espace de stockage.

De façon schématique, nous pouvons considérer que l'analyse comprend les phases suivantes :

- Unifier l'encodage de texte soit pour le corpus, soit pour la requête;
- Normaliser le corpus de textes et les requêtes;
- Découper ou segmenter le texte d'entrée en séquences d'unités lexicales (mot);
- Éliminer les mots vides;
- Déterminer pour chaque mot ses caractéristiques morphologiques;
- Lemmatiser les préfixes et les suffixes, en se basant sur les caractéristiques morphologiques et sur des différents dictionnaires;
- Déterminer les racines possibles pour chaque mot, en se basant sur les dictionnaires de modèles (AOUZANE) et de racines;
- Pondérer les termes générés;
- Créer la base d'index ;

L'indexation (texte en entrée) et la recherche (les requêtes de l'utilisateur) sont traitées par ces modules pour obtenir des résultats pertinents, donc améliorer la recherche (Figure 5.4).

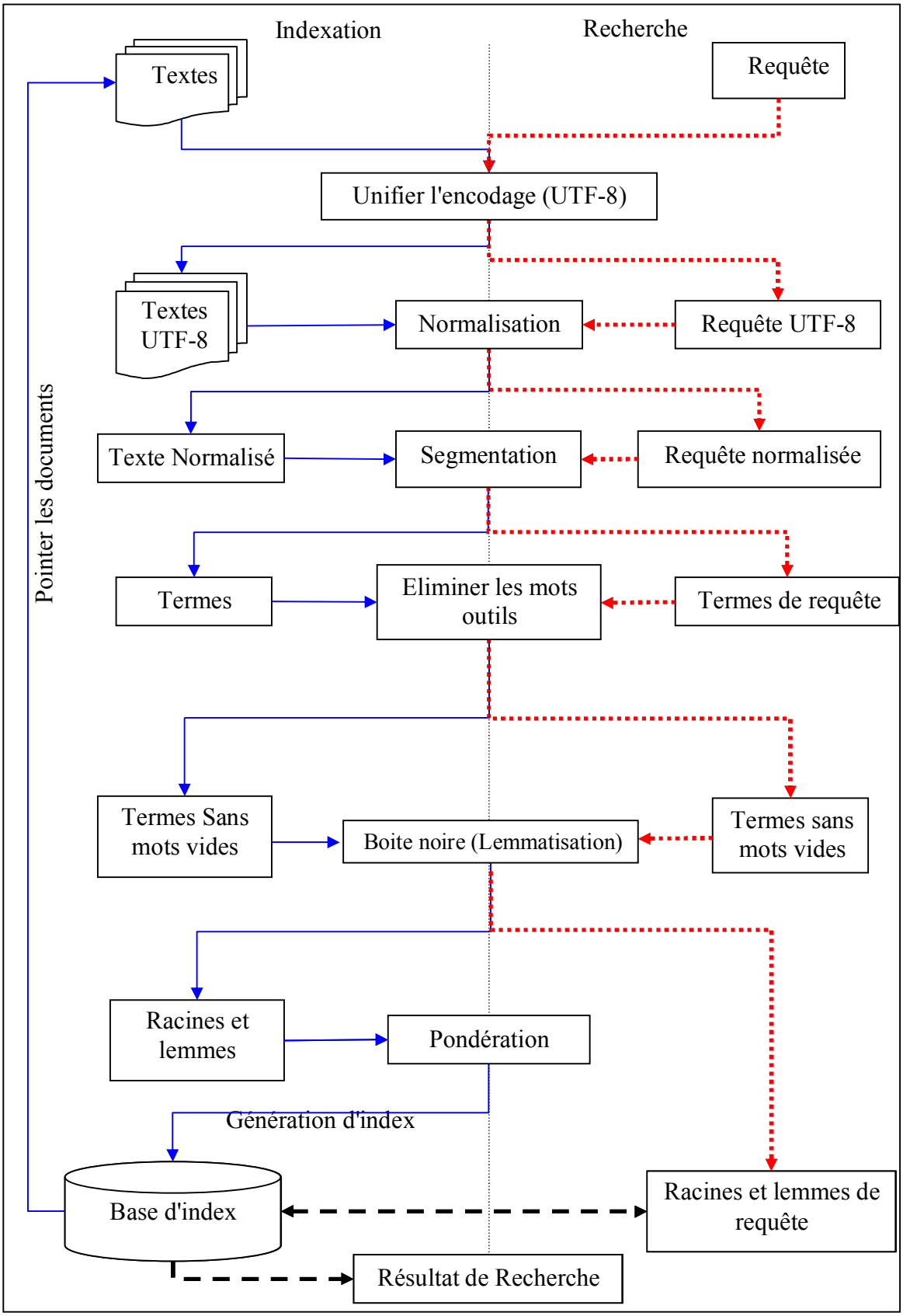

Fig.5.4 : Architecture de Notre Système.

3.3.1 Encodage

La collection de textes et les requêtes peuvent être encodées différemment, les rendant incomparables. Par exemple, les documents sont représentés en Unicode (UTF-8) et les requêtes

en ISO-8859-6 ou un autre encodage. Afin d'unifier les documents avec les requêtes, nous devons réutiliser des outils de conversion entre différents encodages. Ainsi, tout serait transformé en format Unicode dans notre cas.

3.3.2. Normalisation

Comme nous l'avons déjà précisé dans le troisième chapitre, pour manipuler les variations du texte qui peuvent être représentées en arabe, nous avons exécuté plusieurs genres de normalisation sur le texte de corpus (documents et requêtes).

3.3.3. Segmentation

La segmentation est une étape nécessaire et signifiante dans le traitement du langage naturel (voir chapitre 3). La fonction d'un segmenteur est de couper un texte courant en segments de sorte qu'ils puissent être introduits dans un capteur morphologique ou un étiqueteur de position. Le segmenteur est responsable dans un premier temps de définir des limites de mot, celui-ci se fonde principalement sur les espaces blancs et les signes de ponctuation comme des séparateurs entre les mots ou des *segments principaux* (Figure 5.5).

Nous avons développé un segmenteur qui répond à toutes les fonctions citées précédemment.

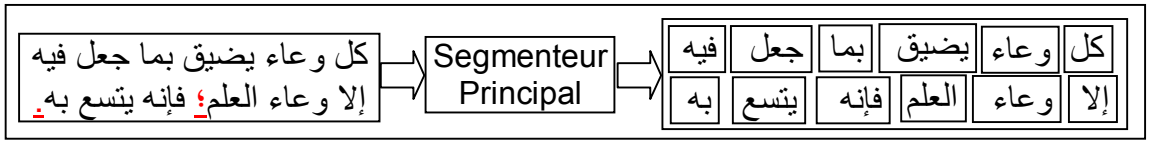

Fig.5.5 : Exemple d'un Segmenteur.

3.3.4. Élimination des mots vides

Un des problèmes majeurs de l'indexation consiste à extraire les termes significatifs et à éviter les mots vides. On distingue deux techniques pour éliminer les mots vides :

- L'utilisation d'une liste de mots vides (aussi appelée anti-dictionnaire),
- L'élimination des mots dépassant un certain nombre d'occurrences dans la collection.

Nous avons utilisé la première technique et à l'aide de la deuxième technique nous avons enrichi notre liste des mots vides.

ؤ	انتن	ؤ	ايضا	أيه	بأيها	بين	دونك	عليكما	عنهما
ئ	انه	ئ	اين	أيها	بأيهم	بينما	ذ	عليكن	عني
ا	انها	ا	آ	إ	بأيهما	بينه	ذا	علينا	غ
اذ	انهم	ابان	أ	إحدى	بذلك	بينها	ذات	عليه	ف
اذا	انهما	ابدا	ألا	إذ	بعد	بينهم	ذاته	عليها	فالتي
اذما	اهلا	اتجاه	أما	إذا	بعدك	بينهما	ذاتها	عليهم	فالذي

Tableau 5.2 : un aperçu sur les mots vides.

Même si l'élimination des mots vides a l'avantage de réduire le nombre de termes d'indexation, elle peut cependant réduire le taux de rappel; c'est à dire la proportion de documents pertinents retournés par le système par rapport à l'ensemble des documents pertinents.

3.3.5. Lemmatisation:

Comme nous avons vu, le traitement morphologique est le cœur de la recherche d'information arabe, plus précisément c'est la lemmatisation qui joue un rôle important. Alors nous appliqué cinq méthodes différentes de lemmatisation et nous avons comparé les résultats et adopté la méthode qui donne une meilleure performance dans la recherche d'information.

3.3.5.1. La méthode PS-M :

La méthode PS-M (ou bien Préfixe Suffixe Sans Modèle) repose sur la réduction des mots fléchis en retirant premièrement ses préfixes et en second lieu ses suffixes selon la méthodologie proposées par Kadri [Kad08], et à chaque étape nous vérifions l'existence de mot résultant dans le dictionnaire de racines, s'il existe on doit arrêter le processus, sinon on doit continuer jusqu'au bout. Lorsque ce processus est fait correctement, il devient facile d'extraire les lettres de lemme, par exemple (لنموه, pour sa croissance) si on retire le suffixe d'abord (وه), alors on va perdre le lemme correct (Figure 5.6).

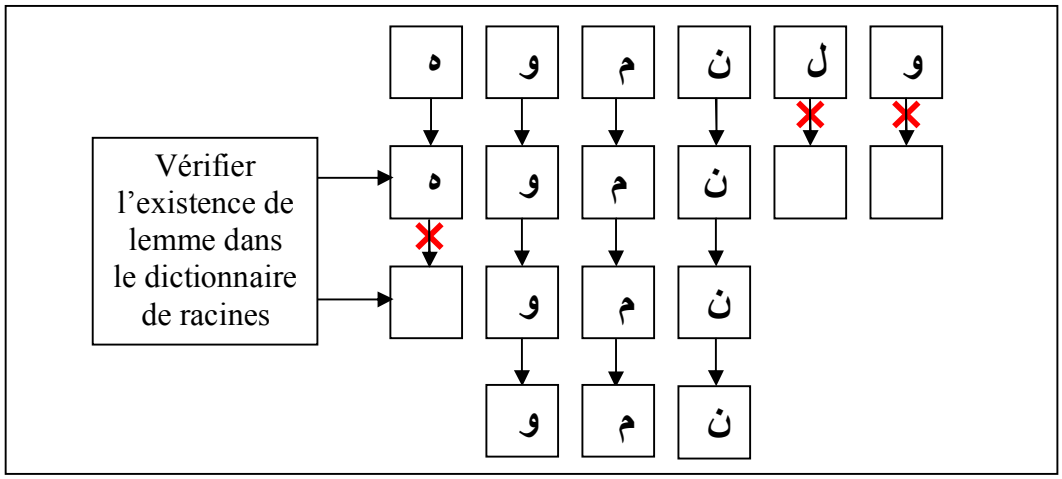

Fig.5.6 : Exemple sur la méthode PS-M.

3.3.5.2. La méthode SP-M :

La méthode SP-M (ou bien Suffixe Préfixe Sans Modèle) repose sur le même principe de la méthode PS-M, mais en retirant premièrement les suffixes et en second lieu les préfixes, par exemple (الالتزامات, les engagements) si on retire le préfixe d'abord (الا), alors on va perdre le lemme correct (Figure 5.7).

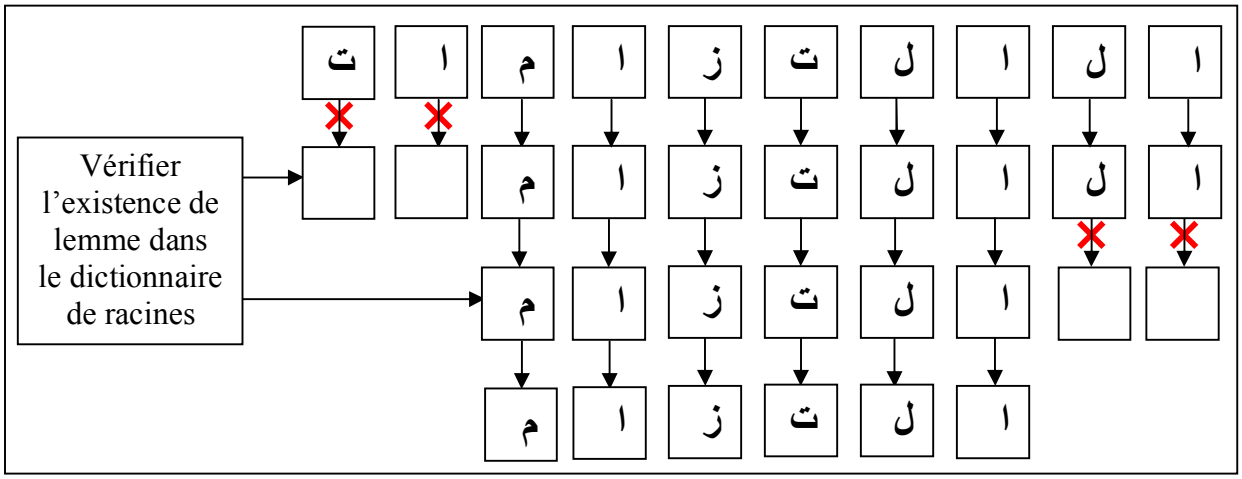

Fig.5.7 : Exemple sur la méthode SP-M.

3.3.5.3. La méthode PS+M (Préfixe Suffixe Avec Modèle):

Après avoir retiré tous les préfixes puis les suffixes du mot fléchi, nous avons comparé celui-ci avec tous les modèles disponibles. Si un modèle est trouvé nous procédons alors à l'extraction des lettres qui forment la racine, si aucun modèle n'est trouvé, nous retournons le mot fléchi tel qu'il est.

Retirer quelques préfixes et suffixes des mots aide à la réduction du nombre des modèles, facilite le processus de correspondance des modèles et permet à plusieurs variations du lemme d'être combinées au même modèle [Als98]. Par exemple nous n'avons conservé aucun de ces modèles: «استفعل», «مستفعل» parce que les deux préfixes « است », « مست » sont existent. Au lieu de cela, nous avons retiré tous ces préfixes et suffixes avant de comparer le mot avec son modèle. Cette manière réduit le nombre des modèles et facilite de trouver le modèle correct.

Nous avons comparé n'importe quel mot avec des modèles selon sa longueur, en utilisant un ensemble de conditions pour vérifier les lettres d'infixe dans le mot. Par exemple, le mot « حواسيب » a la longueur 6, donc nous avons recherché les modèles en utilisant les conditions suivantes :

Trouver un modèle avec la longueur 6 qui a :

- le « و » comme deuxième lettre;
- le « ا » comme troisième lettre;
- le « ي » comme cinquième lettres.

Ces conditions correspondent seulement au modèle « فواعيل ».

En suite, nous avons retiré ces lettres et extrait la racine « حسب » (Figure 5.8).

L'ordre de ces règles et les conditions utilisées pour la comparaison sont des facteurs très importants pour garantir une comparaison correcte.

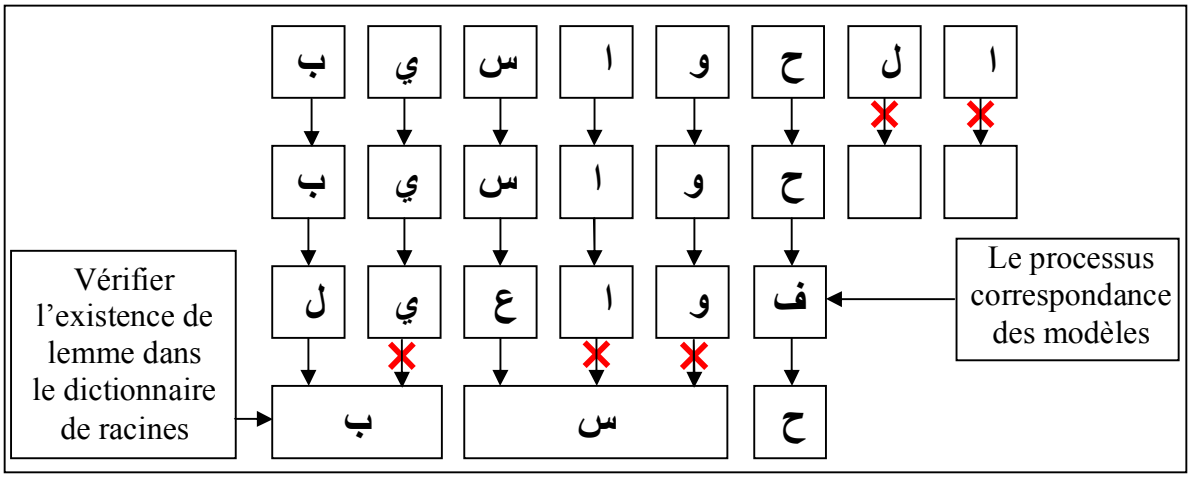

Fig.5.8 : Exemple sur la méthode PS+M.

3.3.5.4. La méthode SP+M :

La méthode SP+M (ou bien Suffixe Préfixe Avec Modèle) repose sur le même principe de la méthode PS+M, mais en retirant premièrement les suffixes et en second lieu les préfixes.

3.3.5.5. La méthode HY (Hybride):

Comme chacune des méthodes de lemmatisation a ses limites, il est naturel de penser à les combiner pour bénéficier des avantages qu'offre chacune d'elles.

A travers la combinaison des techniques de lemmatisation, nous avons amélioré la qualité d'index de corpus (documents et les requêtes), et par conséquent nous avons eu une bonne performance du RI arabe.

3.3.6. Pondération des termes d'indexation

Le calcul de la représentativité d'un terme d'indexation repose sur sa fréquence d'apparition dans le texte en langue naturelle [Sal83]. Afin de mesurer l'importance d'un terme dans un document, nous avons utilisé différentes mesures.

- ***la fréquence relative*** d'un terme d'indexation (***tf***). Il s'agit de la fréquence d'apparition du terme d'indexation dans l'unité documentaire,
- ***la fréquence absolue*** d'un terme d'indexation dans la collection globale d'unités documentaires (***idf***). Il s'agit de la fréquence inverse d'apparition du terme d'indexation dans l'ensemble des unités documentaires de la collection.

Le poids d'un terme d'indexation *i* dans une unité documentaire peut être défini par l'équation suivante [Spa00]:

$$Poids_i = tf_i \,.\, idf_i \qquad (5.1)$$

Où

$$idf_i = \log(N/N_i) + 1 \qquad (5.2)$$

Avec N représentant le nombre d'unités documentaires dans la collection et N_i le nombre d'unités documentaires possédant le terme d'indexation i.

3.3.7. Techniques de création des index

Afin de répondre plus rapidement à une requête, des structures de stockage particulières sont nécessaires pour mémoriser les informations sélectionnées lors du processus d'indexation. Les moyens de stockage les plus répandus sont les suivants : les fichiers inverses (***inverted files***), les tableaux de suffixes (***suffix arrays***) et les fichiers de signatures (***signature files***).

Nous nous sommes basés dans notre implémentation sur les fichiers inverses qui constituent actuellement le meilleur choix possible pour la plupart des applications [Zob98]. Les fichiers inverses sont composés de deux éléments principaux :

- Le vocabulaire, qui est l'ensemble des différents mots du texte ;
- Les occurrences (***posting***) : pour chaque mot, il s'agit de la liste de toutes les positions dans le texte pour lesquelles le mot apparaît (Figure 5.9).

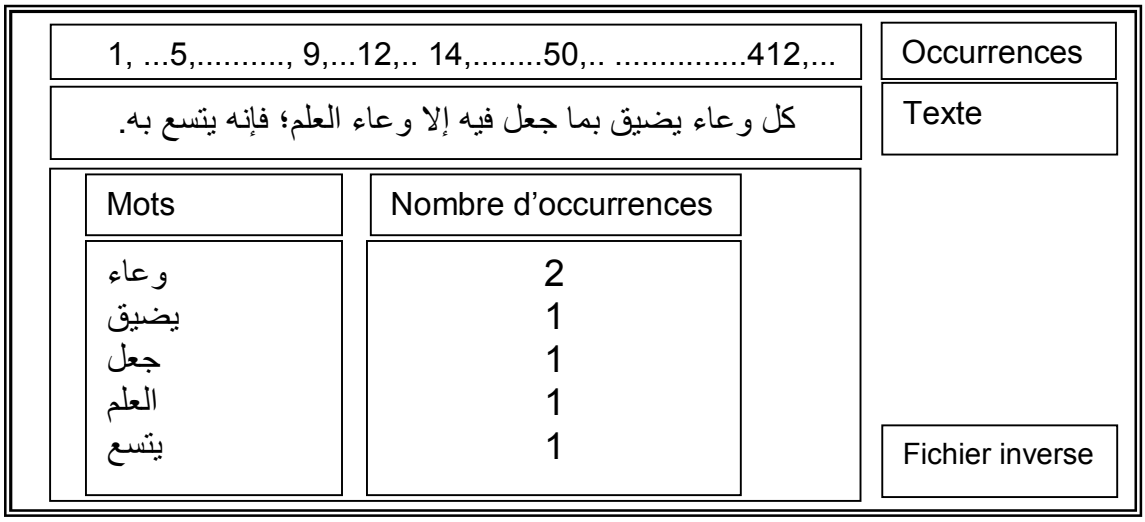

Fig.5.9: Le fichier inverse correspondant à un texte simple.

3.3.8. Méthode de recherche

La requête de l'utilisateur passe par toutes les étapes de l'indexation y compris les étapes de lemmatisation. Les termes de la requête sont mis dans une liste qui sera allégée par les analyses suivantes pour qu'elle soit comparée avec les indexes des documents.

L'utilisateur formule une requête en langue naturelle, le système analyse son contenu et le convertit en éléments du langage d'indexation. Les documents étant représentés par des éléments de ce même langage d'indexation, le système, après comparaison des éléments de la requête avec ceux des documents, détermine les degrés de ressemblance de ces derniers avec la requête et sélectionne ceux qui ont un degré de ressemblance supérieur à un seuil donné.

3.3.8.1. L'appariement document-requête

Avant de décrire le module d'appariement document-requête, rappelons que les documents ne sont pas les seuls à être indexés : les requêtes sont également perçues comme des listes de mots-clés.

La comparaison entre le document et la requête ne permet pas de calculer un score. Cette valeur est calculée à partir d'une fonction ou d'une probabilité de similarité notée ***RSV(Q,d)*** (Retrieval Status Value), où ***Q*** est une requête et ***d*** un document.

La fonction d'appariement est très étroitement liée aux opérations d'indexation et de pondération des termes de la requête et des documents du corpus. D'une façon générale, l'appariement document-requête et le modèle d'indexation permettent de caractériser et d'identifier un modèle de recherche d'information.

La fonction de similarité permet ensuite de classer les documents retournés à l'utilisateur. En effet, l'utilisateur se contente généralement d'examiner les premiers documents retournés. Si les documents recherchés ne sont pas présents dans l'ensemble des premiers documents retournés, l'utilisateur considérera ce système comme mauvais vis-à-vis de sa requête.

4. Expérimentation et évaluation

La lemmatisation est nécessaire pour la performance de RI, elle permet de fusionner les termes ayant un sens similaire avec de petites différences sur la forme morphologique en un seul index, et par conséquent elle permet d'améliorer la qualité de la recherche [Kad08].

Le but de nos expérimentations est d'évaluer les différentes méthodes de lemmatisation sur la performance de recherche d'information arabe.

Une série d'expérimentations a été menée sur notre corpus pour montrer l'effet de chaque méthode de lemmatisation sur la performance de la recherche.

Dans nos expériences, nous avons utilisé les mesures classiques de recherche d'information: précision et rappel. Le tableau 5.2 présente un exemple montrant les résultats des expériences associés à une requête (حرائق النفط).

Nbre doc de corpus	Nbre doc pertinent	Nbre docs pertinents retrouvés					Rappel					Précision				
		PS-M	SP-M	PS+M	SP+M	HY	PS-M	SP-M	PS+M	SP+M	HY	PS-M	SP-M	PS+M	SP+M	HY
3	11	2	2	3	2	3	0.18	0.18	0.27	0.18	0.27	0.67	0.67	1.00	0.67	1.00
6	11	5	5	5	5	5	0.45	0.45	0.45	0.45	0.45	0.83	0.83	0.83	0.83	0.83
9	11	6	8	8	5	7	0.55	0.55	0.73	0.45	0.64	0.67	0.89	0.89	0.56	0.78
12	11	6	9	8	7	9	0.55	0.55	0.73	0.64	0.82	0.50	0.75	0.67	0.58	0.75
15	11	7	10	8	8	10	0.64	0.64	0.73	0.73	0.91	0.47	0.67	0.53	0.53	0.67
18	11	7	11	9	8	11	0.64	0.64	0.82	0.73	1.00	0.39	0.61	0.50	0.44	0.61
21	11	7	11	9	8	11	0.64	0.64	0.82	0.73	1.00	0.33	0.52	0.43	0.38	0.52
24	11	9	11	9	10	11	0.82	0.82	0.82	0.91	1.00	0.38	0.46	0.38	0.42	0.46
27	11	9	11	9	10	11	0.82	0.82	0.82	0.91	1.00	0.33	0.41	0.33	0.37	0.41
30	11	9	11	11	10	11	0.82	0.82	1.00	0.91	1.00	0.30	0.37	0.37	0.33	0.37

Tableau 5.3 : Un exemple sur les résultats des expériences « حرائق النفط »

Nous comparons tout d'abord les deux méthodes de lemmatisation (PS-M et SP-M) que nous avons proposées. La figure 5.10 dresse une comparaison entre ces deux méthodes en fonction de leurs courbes rappel-précision. Les résultats montrent que la méthode de lemmatisation SP-M est uniformément plus efficace que la méthode PS-M presque sur tous les points de rappel ; la courbe SP-M représentant la précision de recherche en fonction des points de rappel est toujours au dessus de la courbe PS-M.

Ces résultats prouvent qu'une lemmatisation PS-M pour les mots arabes n'est pas la meilleure méthode pour la RI arabe. Alors que, la méthode SP-M peut mieux déterminer le noyau sémantique d'un mot et par conséquent elle augmente la performance de la RI.

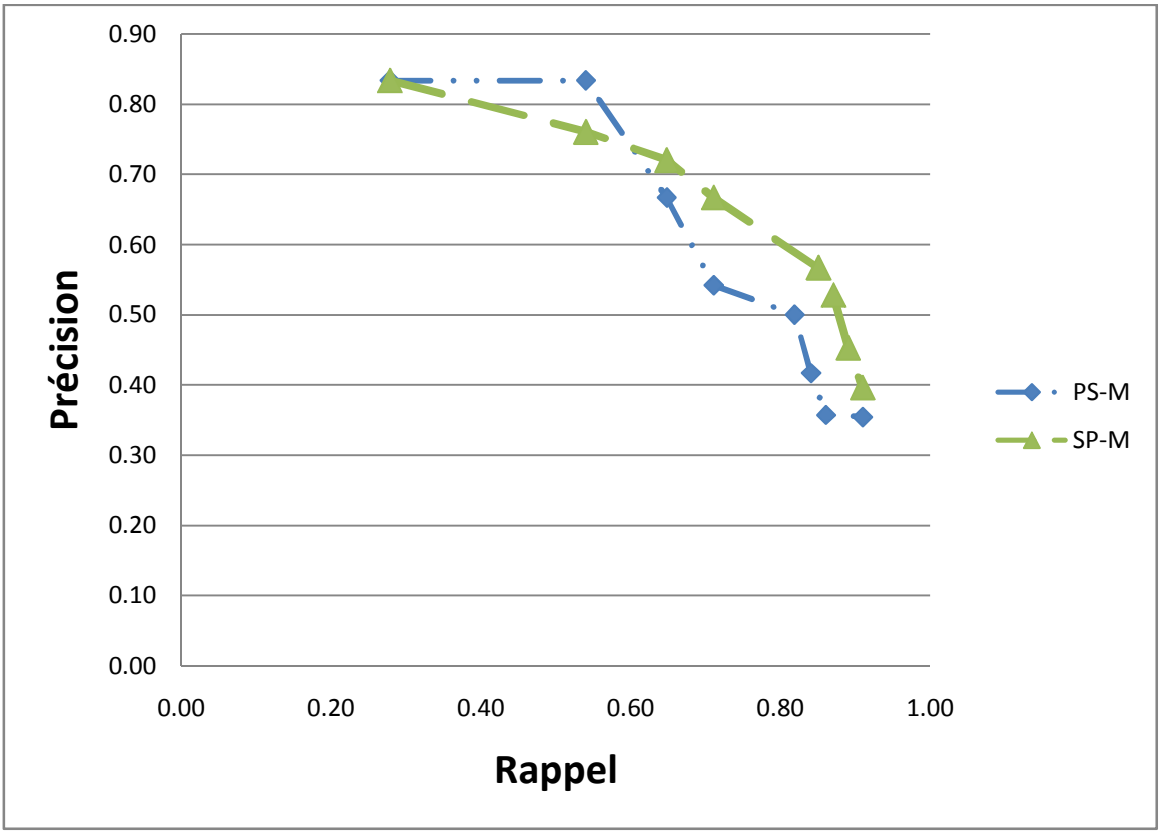

Fig.5.10: Les courbes rappel-précision des deux méthodes de lemmatisation PS-M et SP-M.

Nous comparons par la suite les deux méthodes de lemmatisation (PS+M et SP+M), parce que la lemmatisation à base de ces deux méthodes procède différemment, elle introduit un nouveau facteur, c'est le modèle (OUAZENE), produit en conséquence un ensemble de lemmes candidats, et en utilisant le dictionnaire des racines pour choisit le meilleur lemme.

Afin de pouvoir comparer la méthode PS+M avec la méthode SP+M, nous avons tracé la courbe Rappel-Précision. La figure 5.11 dresse une comparaison entre ces deux méthodes de lemmatisation.

Contrairement à la lemmatisation sans modèle, les résultats montrent que la méthode de lemmatisation PS+M est plus efficace que la méthode SP+M sur tous les points de rappel ; la

courbe PS+M représentant la précision de recherche en fonction des points de rappel est toujours au dessus de la courbe SP+M.

Ces résultats prouvent que la méthode PS+M peut mieux déterminer le noyau sémantique d'un mot, et par conséquent elle augmente la performance de la RI.

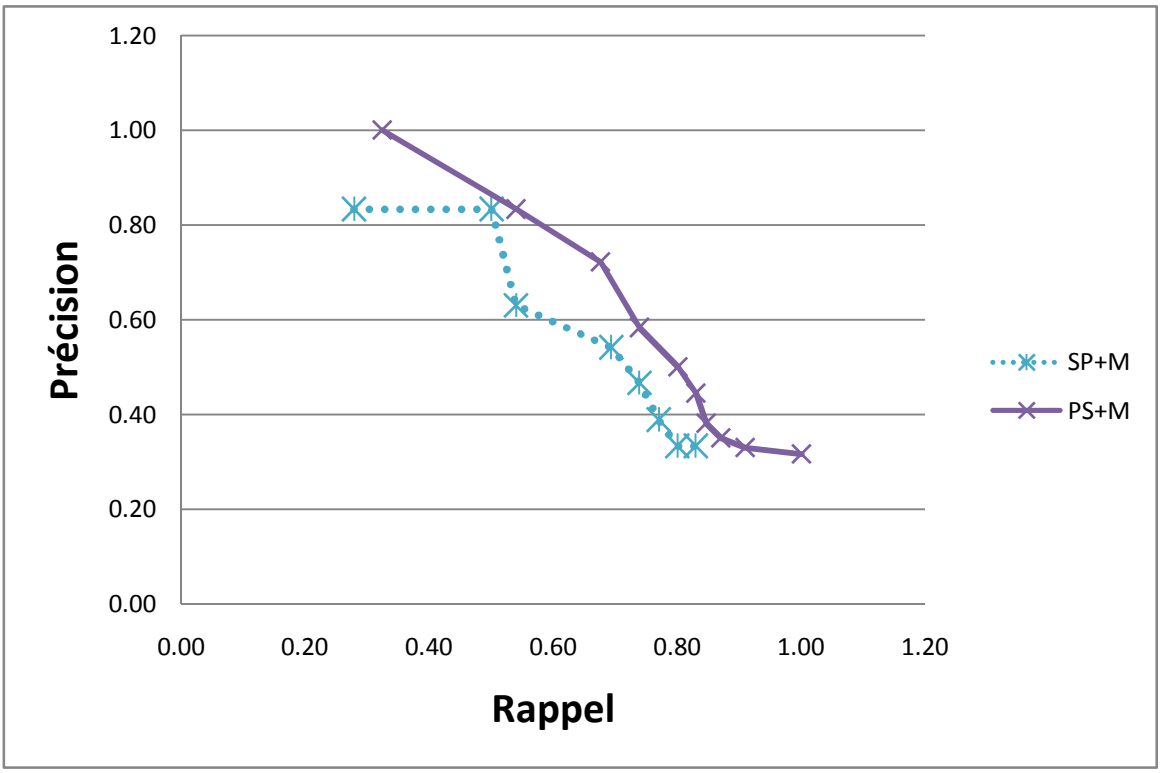

Fig.5.11: Les courbes rappel-précision des deux méthodes de lemmatisation PS+M et SP+M.

La figure 5.12 dresse une comparaison entre les méthodes de lemmatisation sans modèle et avec modèle sur notre collection en fonction de leurs courbes rappel-précision.

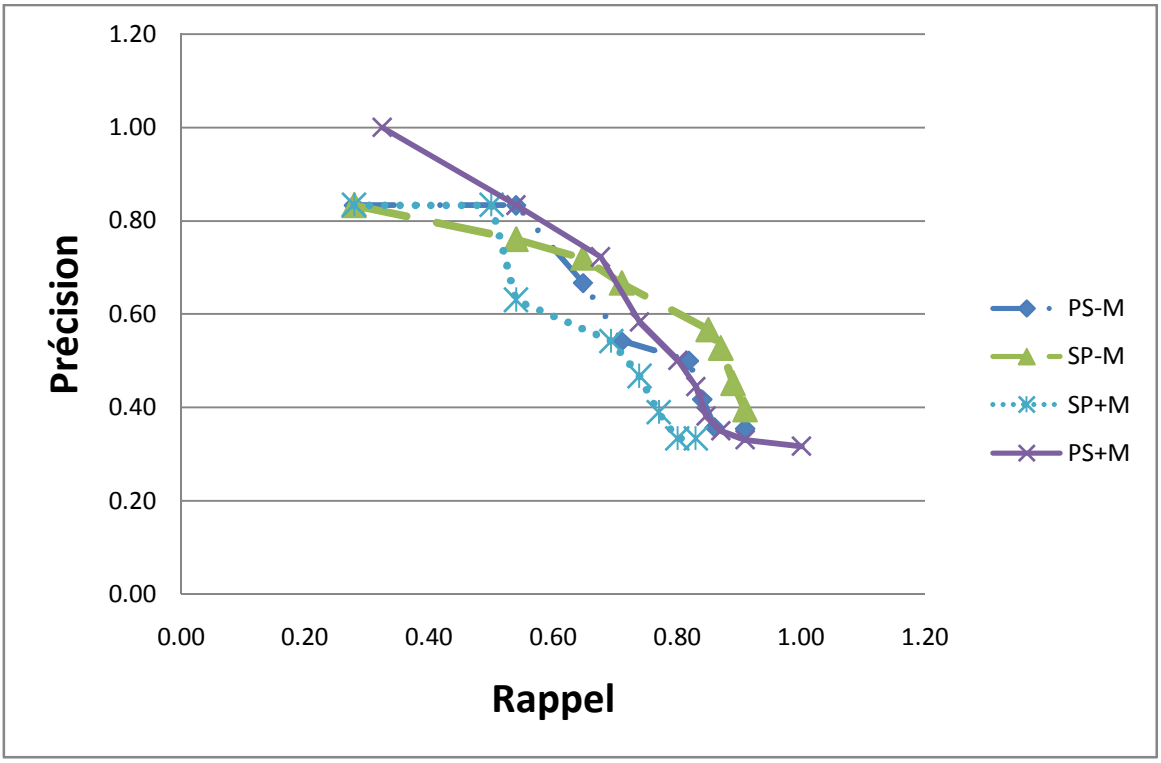

Fig.5.12: Les courbes rappel-précision des méthodes de lemmatisation PS-M, SP-M, PS+M et SP+M.

Sur notre collection, les résultats montrent que la méthode de lemmatisation SP-M est plus efficace que les autres méthodes (PS-M, PS+M, et SP+M). On peut observer ce comportement dans la figure 5.12 : la courbe de lemmatisation SP-M représentant la précision de recherche en fonction des points de rappel est souvent au dessus des autres courbes. Sur l'ensemble des 10 requêtes, nous avons obtenu 57% de précision moyenne avec la méthode de lemmatisation SP-M contre 48%, 51%, et 54% pour les méthodes SP+M, PS-M, et PS+M respectivement.

Cependant, cette figure 5.12 montre que la méthode de lemmatisation PS+M obtient les meilleurs scores quand le rappel est inférieur à 50 %.

Pour cette raison, nous avons proposé une nouvelle méthode de lemmatisation hybride (HY), qui combine toutes les méthodes mentionnées précédemment, et afin d'améliorer la performance globale du processus de lemmatisation. La figure 5.13 dresse une comparaison entre les cinq méthodes de lemmatisation sur notre collection en fonction de leurs courbes rappel-précision.

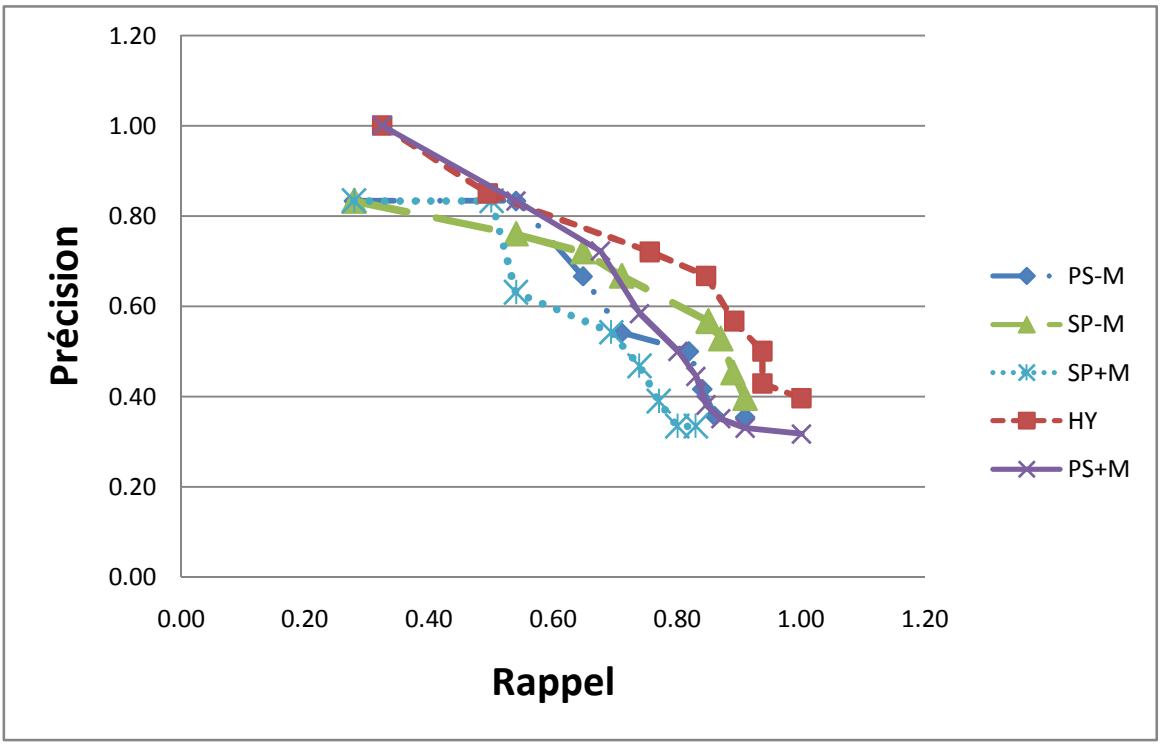

Fig.5.13: Les courbes rappel-précision des cinq méthodes de lemmatisation.

Ces résultats montrent que les deux méthodes de lemmatisation SP-M et HY sont plus efficaces que les autres méthodes (PS-M, PS+M, et SP+M). On peut observer ce comportement dans la figure 5.13 : les courbes de lemmatisation SP-M et HY représentants la précision de recherche en fonction des points de rappel sont souvent au dessus des autres courbes. Nous avons obtenu 58% de précision moyenne avec la méthode de lemmatisation HY et 57% avec la méthode SP-M contre 48%, 51%, et 54% pour les méthodes SP+M, PS-M, et PS+M respectivement.

Cependant, une question posée est quelle est la meilleur méthode SP-M ou HY ? Pour répondre à cette question nous avons mené une autre série d'expérimentations sur notre corpus avec de nouvelles requêtes plus complexes.

Le tableau 5. 3 présente un exemple sur les résultats des expériences qui aient fait avec la requête (المحروقات).

Nbre DocS	Nbre doc pertinent	Nbre docs pertinents retrouvés		Rappel		Précision	
		SP-M	HY	SP-M	HY	SP-M	HY
1	7	1	1	0.14	0.14	1.00	1.00
2	7	2	2	0.29	0.29	1.00	1.00
3	7	2	3	0.29	0.43	0.67	1.00
4	7	2	3	0.29	0.43	0.50	0.75
5	7	2	4	0.29	0.57	0.40	0.80
6	7	2	4	0.29	0.57	0.33	0.67
7	7	2	4	0.29	0.57	0.29	0.57

Tableau 5.4 : Un exemple sur les résultats des expériences « المحروقات »

On remarque une amélioration en précision moyenne de 71% pour une recherche utilisant la méthode HY et de 59% pour une recherche utilisant la méthode SP-M sur notre corpus.

Aussi, on remarque que la méthode de lemmatisation HY est plus efficace que la méthode SP-M. On peut observer cette efficacité dans la figure 5.13 où la courbe de lemmatisation HY représentant la précision de recherche en fonction des points de rappel et souvent au dessus de la courbe SP-M.

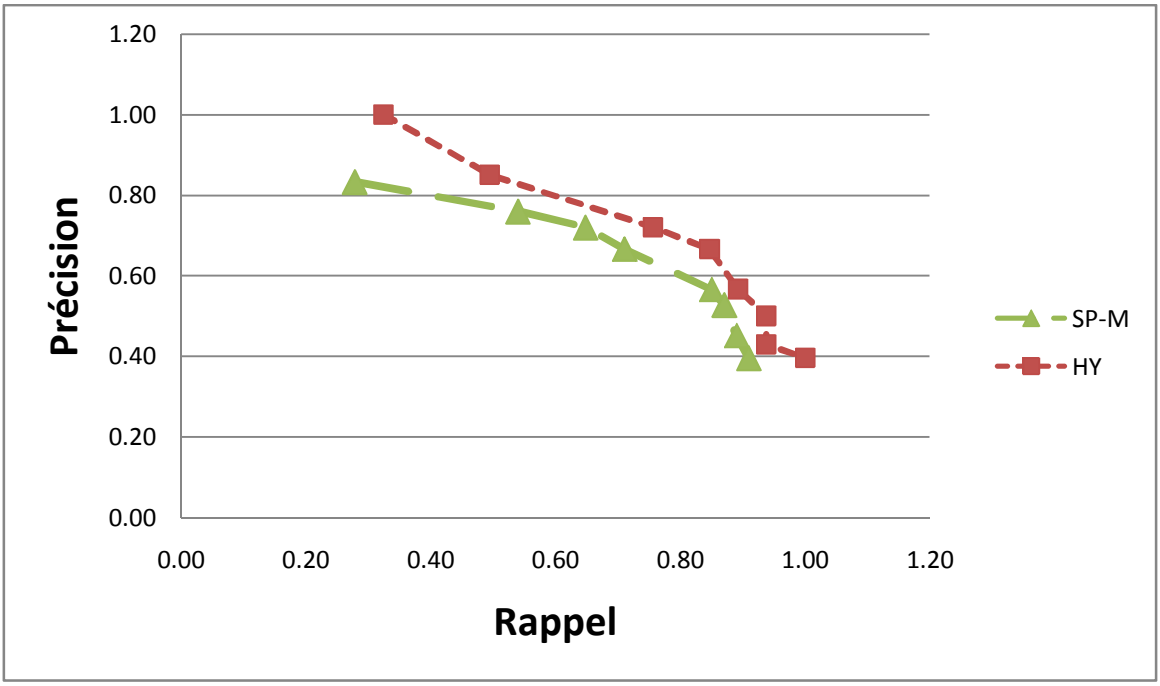

Fig.5.13 : Les courbes rappel-précision des deux méthodes SP-M et HY.

Ces résultats prouvent que la méthode de lemmatisation HY est la meilleure approche, car elle permet avec plus de réussite de grouper beaucoup de mots sémantiquement similaires dans le même index. Cependant, la méthode SP-M ne peut pas déterminer mieux le noyau sémantique d'un mot.

La méthode de lemmatisation HY ne fait pas une troncature aveugle; elle applique différentes décompositions sur le mot original. En cas de présence d'affixes multiples dans un mot, elle permet avec réussite de choisir quel affixe (préfixe ou suffixe) à éliminer d'abord; elle détermine correctement le modèle adéquat, produit en conséquence un ensemble de lemmes candidats en utilisant le dictionnaire des racines, et choisit le meilleur lemme.

La méthode de lemmatisation HY n'est pas parfaite et ne parvient pas à identifier les lemmes corrects pour certains mots ambigus. D'ailleurs c'est dans cet aspect que notre méthode doit être améliorée.

5. Conclusion

L'arabe est l'une parmi les langues les plus utilisées dans le monde, mais relativement, il n'y a que peu d'études qui sont faites sur la recherche d'information et la classification des documents arabes.

L'objectif principal de ce chapitre est d'implémenter un système RI sur les documents textuels arabe, d'expérimenter quelques méthodes de lemmatisation et d'évaluer ces méthodes.

Plusieurs méthodes sont largement investies sur un nombre de traitement de textes et de la recherche d'information.

Le problème principal de chaque méthode proposée est comment identifier les meilleurs termes d'index pour avoir des performances raisonnables ?

Dans ce cadre, nous avons appliqués cinq méthodes différentes de lemmatisation pour résoudre le problème de la performance des systèmes de recherche d'information arabes, et nous avons comparé les résultats et conclu à propos de la méthode qui donne une meilleure performance dans la recherche d'information.

Plus particulièrement, de ces cinq méthodes de lemmatisation, nous avons proposé une nouvelle méthode de lemmatisation hybride (HY), avec cette méthode nous avons essayé de déterminer le noyau d'un mot selon l'intégration de trois techniques différentes (suppression d'affixe, dictionnaires et analyse morphologique) afin améliorer la performance globale du processus de lemmatisation.

La nouvelle méthode présente une meilleure performance de recherche que les autres méthodes, parce qu'elle permet de mieux déterminer le lemme d'un mot, Alors que les autres méthodes ne permettent pas avec réussite de grouper beaucoup de mots sémantiquement similaires dans le même index.

Cependant, la nouvelle méthode peut également entraîner des erreurs à cause de l'ambiguïté. Quelques cas d'ambiguïté présents dans cette méthode, ne posent pas de problèmes pour la recherche d'information, parce que ces mêmes mots présents dans les textes sont lemmatisés de la même façon et par conséquent leurs lemmes identifiés sont identiques aux lemmes obtenus pour ces mots dans les requêtes.

D'autres cas d’erreurs apparaissent parfois quand des termes qui ne sont pas sémantiquement semblables sont groupés dans une classe d'équivalence. C'est d'ailleurs dans cet aspect que notre méthode doit être améliorée.

CONCLUSIONS ET PERSPECTIVES

1. Conclusion

Le problème principal de la recherche d'information (RI) est de trouver les documents qui satisfont les besoins d'un utilisateur (habituellement exprimés sous forme d'une requête) en termes d'informations. Afin d'atteindre cet objectif, une comparaison entre les mots contenus dans la requête et ceux représentant le contenu de chaque document doit être faite. En d'autres termes, un système de recherche d'information doit représenter, stocker et organiser l'information puis fournir à l'utilisateur les éléments correspondant au besoin d'information exprimé par sa requête. Notre travail s'inscrit dans le cadre de l'indexation et de la recherche d'information dans la langue arabe. Ainsi un système de recherche d'information en langue arabe doit prendre en considération ces caractéristiques et proposer des outils et des techniques afin de permettre son traitement automatique. L'objectif général de ce travail est d'implémenter un outil d'indexation et de recherche des textes écrit en langue arabe.

Nous avons étudié certaines caractéristiques de la langue arabe, notamment celles d'ordre morphologique. Contrairement aux autres langues, la langue arabe possède un système dérivationnel très riche et c'est dans cette caractéristique que réside la difficulté de son traitement. Ces caractéristiques constituent en effet les problèmes majeurs face aux travaux sur la langue arabe dans le domaine de la recherche d'information. Nous avons présenté les prétraitements principaux (l'encodage, la segmentation, la suppression des mots vides et la normalisation) pour chaque corpus de texte arabe. Afin de manipuler les variations du texte qui peuvent être représentées en arabe, nous avons exécuté plusieurs genres de normalisation sur le texte dans le corpus et dans les requêtes.

Ayant construit un corpus à partir du web, nous avons effectué une recherche sur le web, et nous avons trouvé la collection « Al-Khat Alakhdar ». Cette dernière à laquelle nous avons intégré quelques requêtes et jugements de pertinences fait l'objet d'une importante production langagière en arabe et contient 694 articles couvrant la période de septembre 2001 jusqu'à ce jour et représentant un volume de 15 MOctets.

Nous nous sommes basés dans notre implémentation d'un système RI sur deux grands axes : l'indexation et la recherche. En effet, la phase d'indexation consiste à construire au préalable une structure d'accès aux documents qui facilitera la phase de la recherche. Celle-ci consiste à retrouver les documents les plus pertinents par rapport à une requête donnée; plus la phase d'indexation est sophistiquée, plus la phase de recherche est facile. En général, les documents retournés sont ordonnés à l'aide d'une mesure de similarité calculée entre le document et la requête.

Pendant la phase d'indexation, le traitement morphologique et plus particulièrement la lemmatisation joue un rôle important. Quoique la lemmatisation fasse l'objet de plusieurs travaux, cet aspect n'est pas encore largement étudié et il semble que les idées ne sont pas encore claires à propos de quel type de lemmatisation est approprié pour la RI arabe.

Dans cette optique, nous avons défini une méthodologie pour améliorer la performance de la recherche des documents arabes. Nous avons appliqué cinq méthodes différentes de lemmatisation, et avons comparé les résultats. Ainsi nous sommes arrivés à en déduire la méthode qui donne une meilleure performance dans la recherche d'information. En effet, parmi ces cinq méthodes de lemmatisation, nous avons proposé une nouvelle méthode de lemmatisation hybride (HY); avec cette méthode nous avons essayé de déterminer le noyau d'un mot selon l'intégration de trois techniques différentes (suppression d'affixe, dictionnaires, et analyse morphologique) pour améliorer la performance globale du processus de lemmatisation. Cette nouvelle méthode présente une meilleure performance de recherche que les autres méthodes, parce qu'elle peut mieux déterminer le lemme d'un mot. Ainsi nous avons montré l'efficacité de l'enlèvement des suffixes avant les préfixes pendant l'étape de lemmatisation pour les textes arabes.

Cependant, cette méthode peut également entraîner des erreurs, et ce quand des termes se regroupent dans une classe d'équivalence malgré qu'ils ne soient pas sémantiquement semblables, car un tel lemmatiseur peut grouper des termes d'une manière erronée. C'est d'ailleurs dans cet aspect que notre méthode doit être améliorée.

2. Perspectives

Les différentes pistes explorées pendant ce travail nous ont amenées à envisager de nombreuses perspectives. Nous présentons ici celles qui nous paraissent les plus prometteuses.

2.1. Lemmatisation des mots

Notre analyseur morphologique a ses limites et entraîne parfois des erreurs de lemmatisation. Il est très difficile de développer un analyseur morphologique parfait en arabe à cause de la complexité de la structure linguistique arabe et plus particulièrement les mots ambigus. Quand plusieurs lemmes candidats sont proposés pour un mot, plus de traitement au niveau linguistiques et statistiques de corpus doivent être faits pour choisir le meilleur lemme. Par exemple, on peut utiliser un dictionnaire pour déterminer le lemme approprié.

2.2. Approche Sémantique

Le Web sémantique apparaît comme la prochaine génération du Web dont le but est de donner à l'information sur le Web une représentation sémantique afin d'être accessible et compréhensible par des machines.

L'intérêt d'utiliser des ressources sémantiques en recherche d'information est de pouvoir retourner, lors d'une recherche par similarité, les documents qui partagent avec la requête le maximum de concepts plutôt que le maximum de mots-clés, et par conséquent augmenter la précision.

2.3. Approche Hybride

Comme chacune des approches ont leurs propres limites, il est normal de penser à combiner les approches pour bénéficier des avantages qu'offrent chacune d'entre elles, ce qui mène à l'amélioration de la performance globale du système de recherche d'information.

BIBLIOGRAPHIE

- [Abb04] Abbes R., La conception et la réalisation d'un concordancier électronique pour l'arabe, thèse doctorat, L'institut national des sciences appliquées de Lyon, 2004.
- [Abd04] Ahmed Abdelali. Improving Arabic Information Retrieval Using Local variations in Modern Standard Arabic, New Mexico Institute of Mining and Technology, 2004.
- [AFN93] AFNOR, Information et documentation. Principes généraux pour l'indexation des documents, NF Z 47-102, 1993.
- [Ale02] Alemayehu, Nega and Willet, P. (2002). Stemming of Amharic words for Information Retrieval. In Literary and Linguistic Computing. 17 (1): 1-17.
- [Alk94] Al-Kharashi, I. and Evens, M. W. Comparing words, stems, and roots as index terms in an Arabic information retrieval system. JASIS, 45 (8), pp. 548-560, 1994.
- [Als98] Al-Shalabi, R. and M. Evens, 1998. A computational morphology system for Arabic. Proceeding of the Workshop on Semitic Languges. COLING-ACL 98, 1998, pp: 58-65.
- [Att06] Attia, 2006 An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks, The Challenge of Arabic for NLP/MT Conference. The British Computer Society, London, UK.
- [Ben06] Benzineb K., Construire un moteur d'indexation, Université de Genève, 2006.
- [Ber01] Berlian, V., Vega, S. N., and Bressan, S. Indexing the Indonesian web: Language identification and miscellaneous issues. Presented at Tenth International World Wide Web Conference, Hong Kong, 2001.
- [Bes01] Beesley, Kenneth R. 2001. Finite-state morphological analysis and generation of Arabic at Xerox Research: Status and plans in 2001. In ACL Workshop on Arabic Language Processing: Status and Perspective, pp. 1-8, Toulouse, France, July 6th.
- [Bes03] Kenneth R. Beesley and Lauri Karttunen, Finite State Morphology, CSLI Publications, 2003.
- [Car01] Carlberger, J., Dalianis, H., Hassel, M., and Knutsson, O. Improving precision in information retrieval for Swedish using stemming. In Proceedings of NODALIDA '01 - 13th Nordic conference on computational linguistics. Uppsala, Sweden, 2001.
- [Cha96] Chanod J-P, Tapanainen P (1994): A Non-Deterministic Tokenizer for Finite-State Parsing, ECAI'96. Budapest, Hungary.
- [Che02] Chen, A., and Gey, F. Building an Arabic stemmer for information retrieval. In TREC 2002. Gaithersburg: NIST, pp 631-639, 2002.
- [Coo88] Cooper, W. (1988) «Getting Beyond Boole," Information Processing & Management.

- [Dar02] Darwish, K. and Oard, D.W. CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval. In TREC 2002. Gaithersburg: NIST, pp 703-710, 2002.

- [Dia04] Diab M, Hacioglu K, Jurafsky D (2004): Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks, Proceedings of NAACL-HLT 2004. Boston.

- [Dic01] On lemmatization in Arabic. A formal definition of the Arabic entries of multilingual lexical databases, ACL 39th Annual Meeting. Workshop on Arabic Language Processing; Status and Prospect. Toulouse, pp 23-30.

- [Ekm96] Ekmekcioglu, F. C., Lynch, M. F., and Willett, P. Stemming and n-gram matching for term conflation in Turkish texts. Information Research News, 7 (1), pp. 2-6, 1996.

- [Fox88] Fox, E. et Koll, M. (1988). Fox, E. & Koll, M. (1988); Partial Enhanced Boolean Retrieval: Experiments with the SMART and SIRE Systems», Information Processing & Management.

- [Fra92] Frakes et Baeza-Yates (1992). «Information Retrieval: Data Structures & Algorithms». Prentice Hall.

- [Gar01] Carlberger, J., Dalianis, H., Hassel, M., and Knutsson, O. Improving precision in information retrieval for Swedish using stemming. In Proceedings of NODALIDA '01 - 13th Nordic conference on computational linguistics. Uppsala, Sweden, 2001.

- [Gey01] Gey, Fredric C. Oard, Douglas. 2001. The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries. TREC 2001. pp. 16-25

- [Gim01] Girma Berhe. (2001). A Stemming Algorithm Development for Tigrigna Language Text Documents.

- [Gre96] Greengrass, M., Robertson, A. M., Robyn, S., and Willett, P. Processing morphological variants in searches of Latin text. Information research news, 6 (4), pp. 2-5, 1996.

- [Hab05] Habash, Rambow, Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In ACL'05, Ann Arbor, MI, USA.

- [Hul96] Hull, D. Stemming Algorithms - A Case Study for Detailed Evaluation. JASIS, 47(1):70-84, 1996.

- [Kad08] Kadri, Y. 2008. Recherche d'Information Translinguistique sur les Documents en Arabe. Thèse présentée à la Faculté des études supérieures en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.) en informatique. Septembre 2008.

- [Kho99] Khoja, S. Stemming Arabic Text. Lancaster, U.K., Computing Department, Lancaster University. www.comp.lancs.uk/ computing/users/khoja/stemmer.ps, 1999.

- [Kra96] Kraaij, W. and Pohlmann, R. Viewing stemming as recall enhancement. In Proceedings of ACM SIGIR96. pp. 40-48, 1996.

- [Kro93] Krovetz, R. Viewing morphology as an inference process. In Proceedings of ACM-SIGIR, pp. 191–202, 1993

- [Lan93] Lancaster and Warner (1993). «Information Retrieval Today». Information Resources, Arlington VA, 1993.

- [Lar01] L. S. Larkey and M. E. Connell. Arabic information retrieval at UMass in TREC-10. In Proceedings of the Text REtrieval Conference (TREC-10), pages 562 570, Gaithersburg, Maryland, 2001.

- [Lar02b] Larkey, Leah S. and Connell, Margaret. 2002. Arabic Information Retrieval at UMass in TREC-10. The Tenth Text Retrieval Conference. NIST, pp. 562-570.

- [Lar03] Larkey, Leah S., James Allan, Margaret E. Connell, Alvaro Bolivar, and Courtney Wade. 2003. UMass at TREC 2002: Cross Language and Novelty Tracks. Ellen M. Voorhees and Lori P. Buckland (Eds.) TREC 2002, NIST Special Publication 500-251, pp 721-732.

- [Lar05] Leah S. Larkey, Lisa Ballesteros, Margaret E. Connell. Light Stemming for Arabic Information Retrieval, 2001. The Center for Intelligent Information Retrieval and SPAWARSYSCENSD grant numbers N66001-99-1-8912 and N66001-02-1-8903. 2005.

- [Lov68] Lovins, J. B. Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11, pp. 22- 31, 1968.

- [Man99] Manning & Schütze, 1999 “Foundations of Statistical Natural Language Processing.”

- [Mar91] Marcus, 1991 «Computer and Human Understanding in Intelligent Retrieval Assistance», American Society for Information Science, 28, 1991.

- [McN02] P. McNamee, C. Piatko, and J. May eld. JHU/APL at TREC 2002: Experiments in Filtering and Arabic Retrieval. In Proceedings of the Text Retrieval Conference (TREC- 11), pages 358 363, Gaithersburg, Maryland, 2002.

- [Mou01] Moulinier, I., McCulloh, A., and Lund, E. West group at CLEF 2000: Non-English monolingual retrieval. In Crosslanguage information retrieval and evaluation: Proceedings of the CLEF 2000 workshop, C. Peters, Ed.: Springer Verlag, pp. 176-187, 2001.

- [Nel05] Nelken and Shieber, 2005, Arabic Diacritization Using Weighted Finite-State Transducers, Proceedings of the 2005 ACL Workshop on Computational Approaches to Semitic Languages. Michigan.

- [Nic92] Nicholas J. Belkin and W. Bruce Croft (1992). «Information Filtering and Information Retrieval: Two Sides of the Same Coin», Communication of the ACM, Dec., 1992.

- [Pir01] Pirkola, A. Morphological typology of languages for IR. Journal of Documentation, 57 (3), pp. 330-348, 2001.

- [Pop92] Popovic, M. and Willett, P. The effectiveness of stemming for natural-language access to Slovene textual data. JASIS, 43 (5), pp. 384-390, 1992.

- [Por80] Porter, M. F. An algorithm for suffix stripping. Program, 14 (3), pp. 130-137, 1980.
- [Sal83] SALTON G., MACGILL M.J., « Introduction to modern information retrieval », McGraw Hill International Book Company, ISBN 0-07- Y66526-5, 1983.
- [Sal90] Salton, G. and Buckley, C. (1990) «Improving Retrieval Performance by Relevance Feedback», Journal of American Society for Information Science.
- [Sch97] P. Schauble, Multimedia Information Retrieval: content-based Information Retrieval from Large Text and Audio Databases, Kluwer Academic Publishers, 1997
- [Spa00] K. Sparck-Jones, S. Walker, and S. Robertson. A probabilistic model for information retrieval/development and comparative experiments, part 1 and 2. Information Processing and Management, 36(6) :pages 779–840, 2000
- [Spo93] Spoerri, A. (1993) «Visual Tools for Information Retrieval», Proc. IEEE Workshop on Visual Languages.
- [Tai00] Tai, S. Y., Ong, C. S., and Abdullah, N. A. On designing an automated Malaysian stemmer for the Malay language. (poster). In Proceedings of the fifth international workshop on information retrieval with Asian languages, Hong Kong, pp. 207-208, 2000.
- [Tou05] المعرب والدخيل في اللغة العربية وآدابها تأليف د. محمد التونجي دار المعرفة بيروت- لبنان الطبعة الاولي. Année (2005) ISBN 9953-446-75-X.
- [Van79] C. J. VAN RIJSBERGEN. Information Retrieval. Butterworths, USA, 1979.
- [Wig98] Wightwick, J. and Gaafar, M. Arabic verbs and essentials of grammar. Chicago: Passport Books, 1998.
- [Xuj02] Xu, J., Fraser, A., and Weischedel, R. Empirical studies in Strategies for Arabic retrieval. In Sigir 2002. Tampere, Finland: ACM, 2002.
- [Zaj01] Zajac, Rémi, Malki, Ahmed, Abdelali, Ahmed, Cowie, James, Ogden William C. 2001. Arabic-English NLP at CRL, Proceedings of the Arabic NLP Workshop ACL/EACL 2001.
- [Zob98] J. Zobel, A. Moffat, and K. Ramamohanarao. Inverted files versus signature files for text indexing. ACM Transactions On Database Systems, 23(4):453–490, December 1998.

RÉSUMÉ

L'Arabe est une langue fortement flexionnelle qui a une structure morphologique complexe. La recherche d'information sur le texte arabe exige la forme de base du mot (racine ou lemme) pour être la plus pertinente, donc le processus de lemmatisation est nécessaire. La lemmatisation peut être définie comme un processus qui consiste à retirer tous les affixes (préfixes, infixes, ou/et suffixes) des mots pour ramener ces derniers à leurs lemmes ou racines.

La complexité morphologique de la langue arabe rend particulièrement difficile le développement des applications pour le traitement en langue naturelle. Dans les langues sémitiques comme l'arabe, la plupart des lemmes de nom, d'adjectif, et de verbe sont dérivés de quelques mille racines par l'insertion de nouvelles lettres.

Dans ce travail, nous avons proposé une méthode hybride qui incorpore trois techniques différentes pour que la lemmatisation arabe résolve les problèmes liés à chaque technique. Ces trois techniques sont: suppression d'affixe proposée par Kadri [Kad08], dictionnaires, et analyse morphologique. Ces techniques ont besoin d'une certaine adaptation pour être pertinentes pour l'utilisation. Chaque technique est adaptée individuellement pour résoudre les problèmes pratiques liés à elle-même.

La contribution principale de ce travail concerne la démonstration de l'efficacité de la méthode hybride comparée aux autres méthodes, et le choix de l'enlèvement des suffixes avant les préfixes pendant l'opération de lemmatisation Arabe.

Mots clés : Recherche d'information, lemmatisation, langue arabe.

ملخص

علوم اللغة العربية متعددة منها علم الصرف الذي ميزته الأساس كونه معقدا. و فهم الكلمة وتحليلها أو ما يسمى بالتحليل الصرفي (morphological analysis) هو الأساس في تطبيقات استرجاع المعلومات (Information Retrieval) وهي التطبيقات التي تبنى عليها برامج الأرشفة، محركات البحث، محركات قواعد البيانات، تلخيص النصوص ومعالجة اللغات الطبيعية.

أما تطوير تطبيقات معالجة اللغات الطبيعية يعتمد في الأساس على الوصول إلى جذر الكلمة أو أصلها، و الوصول إلى الجذر أو الأصل في اللغة العربية ليس بالأمر الهين، لأن المعروف عن اللغات السامية أنهامعقدة بنيوياً ، لذلك يظل الوصول إلى الجذر نوعا من التحدي بخصوص تطوير تطبيقات معالجة اللغات الطبيعية. خاصة و أن جل الكلمات العربية تشتق من بضعة ألاف جذر، و بعض الطرق لفهم الكلمة تعتمد على إرجاع الكلمة إلى جذرها (Root) ، بينما يعتمد البعض الآخر على إرجاع إلى كلمة لأصلها (Stem) أي حذف أحرف الزيادة في أول الكلمة و آخرها.

و قد اقترحنا في هذا العمل طريقة هجينة من ثلاث تقنيات مختلفة، للاستفادة من إيجابيات كل تقنية في التعامل مع صعوبات اللغة العربية، هذه التقنيات الثلاث هي : تقنية حذف حروف الزيادة التي اقترحها الباحث قادري [Kad08] ، تقنية القواميس ، و تقنية التحليل الصرفي. و تتطلب هذه التقنيات بعض التعديلات لتصبح جاهزة للاستخدام، كذلك تم تصميم كل تقنية على حدا من أجل حل المشاكل العملية المتعلقة بكل منها.

و قد جاء هذا العمل ليقدم مساهمة إضافية في إثبات فعالية الطريقة الهجينة المقترحة مقارنة مع الطرق الأخرى، وفي اختيار حذف ما زيد في آخر الكلمة قبل ما زيد في أولها أثناء عملية إرجاع الكلمة إلى أصلها.