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

Université Mostefa Ben Boulaïd – Batna 2
Faculté de mathématiques et d'informatique
Département d'Informatique

THESE
Présentée par
Tahar DILEKH

En vue de l'obtention du diplôme de
Doctorat en Sciences en Informatique
Spécialité : **S**ystème d'**I**nformation et de **C**onnaissance (**SIC**)

Un modèle sémantique pour la recherche d'information en langue arabe

Soutenue publiquement le 14 /02 /2019 devant le jury formé de :

Dr. Hamouma MOUMEN	M.C.A	Président	Université de Batna 2
Dr. Saber BENHARZALLAH	M.C.A	Rapporteur	Université de Batna 2
Pr. Mohammed BENMOHAMMED	Professeur	Examinateur	Université de Constantine 2
Dr. Laid KAHLOUL	M.C.A	Examinateur	Université de Biskra
Dr. Abdelhamid DJEFFAL	M.C.A	Examinateur	Université de Biskra
Dr. Larbi GUEZOULI	M.C.A	Examinateur	Université de Batna 2
Dr. Ali BEHLOUL	M.C.A	Invité	Université de Batna 2

REMERCIEMENTS

Je tiens tout d'abord à exprimer ma profonde gratitude au Docteur Saber BENHARZALLAH, maître de conférence au département d'informatique de l'Université de Batna 2, 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 Ali BEHLOUL, maître de conférence au département d'informatique de l'Université de Batna 2, pour son aide, son soutien, ses conseils.

Je remercie le docteur hamouma MOUMEN 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 Mohammed BENMOHAMMED, Laid KAHLOUL, Abdelhamid DJEFFAL, Larbi GUEZOULI qui ont accepté d'examiner ce travail. Je leur suis pleinement reconnaissant pour leur participation à ce jury.

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 cette thèse de doctorat, sans oublier les fruits de ma vie mes trois beaux enfants, ZEINEB, REDHA et CHAHID.

Je veux, bien évidemment, remercier les personnes avec lesquelles j'ai travaillé durant ce parcours car sans le travail collectif, rien n'était possible : Malleki, Laghrib, Merzoug, Amroussi, Sakoub, Guezouli, Noui, Belloula, et bien d'autres.

DEDICACE

Je dédie ce modeste travail à toute ma famille, mes amis, et à tous ceux, de près ou de loin, m'ont aidé.

ABSTRACT

Information Retrieval Systems (IRS) are designed to facilitate access to stored information. Indexing consists of constructing simplified representations (descriptors) describing the informational content of documents and queries in order to facilitate the search. Traditional information retrieval systems rely on approaches that represent documents (respectively queries) with descriptors extracted from their texts. In such systems, document-query matching is lexical based on the presence (or absence) and frequency of the query words in the document. However, the frequencies of the keywords in a document are not sufficient to identify the relations expressed and to locate the information useful for a user request. How then to introduce "more semantics" in the search for information and the textual search? What semantics?

In this thesis, we will present methods to introduce the semantic aspect in the search for information in Arabic language. The main hypothesis is that the inclusion of conceptual knowledge such as dictionaries, thesauri and ontologies in the information retrieval process can contribute to the resolution of major problems currently encountered in the search for information.

In this work, we propose a semantic information retrieval model in Arabic textual documents. We are studying different lemmatization algorithms that have been developed for the Arabic language and we propose a new algorithm that can help to determine the right lemma before disambiguation if there is a possibility of ambiguity. We also present the fundamental elements that are normally found in the knowledge resources for semantic IRS, and we implement a semantic indexing method for information retrieval where we use the “Contemporary Arabic Dictionary” as a lexical resource for explore the impact of switching from classical indexing to semantic indexing.

Keywords: Semantic information retrieval, stemming, indexing, Word sense disambiguation, Arabic NLP.

RÉSUMÉ

Les systèmes de recherche d'information textuelle (SRI) sont conçus pour faciliter l'accès aux informations stockées et l'une des tâches principales d'un SRI est l'indexation, qui consiste à construire des représentations simplifiées (descripteurs) décrivant le contenu informationnel des documents et requêtes en vue de faciliter la recherche. Les systèmes de recherche d'informations classiques reposent sur des approches qui représentent les documents (respectivement requêtes) par des descripteurs extraits à partir de leurs textes. Or, les fréquences de ces descripteurs ne sont pas suffisantes pour identifier les relations exprimées et localiser les informations utiles pour une requête d'utilisateur. Comment alors introduire « plus de sémantique » dans la recherche d'informations et la fouille textuelle ? Quelle sémantique ?

Dans cette thèse, nous présenterons des méthodes pour introduire l'aspect sémantique dans la recherche d'information en langue Arabe. L'hypothèse principale est que l'inclusion de connaissances telles que les dictionnaires, les thesaurus et les ontologies dans le processus de recherche d'information peut contribuer à la résolution de problèmes majeurs actuellement rencontrés dans la recherche d'information.

Dans ce présent travail, nous proposons un modèle de recherche d'information sémantique dans les documents textuels arabes. Nous étudions des différents algorithmes de lemmatisation qui ont été développés pour la langue Arabe et nous proposons un nouvel algorithme qui peut aider à déterminer le bon lemme avant de faire la désambiguïsation, s'il y aurait une possibilité d'ambiguïté. Nous présentons, également, les éléments fondamentaux qui se trouvent normalement dans les ressources de connaissances pour les SRI sémantique, et nous implémentons une méthode d'indexation sémantique pour la recherche d'information où nous utilisons « le dictionnaire de la langue arabe contemporaine » comme ressource lexicale pour explorer l'impact du passage d'une indexation classique à une indexation sémantique.

Mots clés : Recherche d'information sémantique, lemmatisation, indexation, désambiguïsation sémantique, traitement automatique de la langue arabe.

ملخص

تعد أنظمة استرجاع المعلومات النصية إحدى تطبيقات معالجة اللغات الطبيعية. ويتكون أي نظام استرجاع معلومات تقليدي من مرحلتين أساسيتين، وهما: الفهرسة والبحث؛ فالفهرسة الآلية للوثائق النصية أو التكشيف هي مرحلة مهمة وحاسمة في إنجاح أنظمة استرجاع المعلومات، لكونها عملية اختيار مجموعة من الكلمات أو الجمل الكشفية (الواصفات، المحددات ...) التي تعد ممثلة للمحتوى الموضوعي لكل وثيقة، ووضعها في ملف بطريقة تسهل وتسرع في مرحلة لاحقة عملية البحث معتمدة في ذلك على عدد تكرار الكلمات في المستندات. غير أن ترددات الكلمة ليس كافي لتحديد العلاقات وأنواعها بين الكلمات، وبالتالي تحديد الكلمات المفيدة للمستخدم. والسؤال الذي يطرح هو ما هو السبيل لإضفاء الجانب الدلالي في تطبيقات البحث عن المعلومات ؟ وأي نوع من الدلالات ؟

في هذه الأطروحة ، سنقوم بعرض نموذج يساعد على إضفاء الجانب الدلالي في تطبيقات البحث العربية. والفكرة الرئيسية تدور حول إدراج المعرفة المستخرجة من الموارد المعجمية مثل القواميس والمكانز والأنطولوجيات في عملية استرجاع المعلومات.

في هذا الاطار قمنا في مرحلة أولى بتطوير خوارزمية تساعد على الوصول إلى جذع الصحيح للكلمة العربية قبل إزالة الغموض عنها (إذا كانت غامضة)، معتمدين على معجم " اللغة العربية المعاصرة" كمورد معجمي مهم لاستخراج المعاني وفك الغموض وكذلك الأنواع المختلفة لخوارزميات "ليسك" للانتقال من طريقة الفهرسة الكلاسيكية القديمة التي تعتمد على الكلمات إلى الفهرسة الدلالية التي تعتمد على معاني الكلمات.

الكلمات المفتاحية: أنظمة استرجاع المعلومات الدلالية، التجديع، الفهرسة، التكشيف، إزالة الغموض الدلالي، المعالجة الآلية للغة العربية.

TABLE DES MATIERES

LISTE DES TABLEAUX

LISTE DES FIGURES

Introduction générale

INTRODUCTION GENERALE

1.1. Contexte

De nos jours, l'information joue un rôle primordial dans notre vie quotidienne ; nous avons besoin de l'information pour prendre les meilleures décisions possibles. Dans chacune de nos activités personnelles, les prises de décisions requièrent l'information qui les soutient, l'information étant nécessaire dans presque tous les domaines de la pensée et l'action humaine. En revanche, l'information numérique et les nouvelles formes de technologies d'information sont devenues le centre d'intérêt de notre société. La croissance des technologies d'information (TI) a permis la disponibilité de l'information enregistrée.

Une partie considérable de ces informations disponibles est codé en langue arabe, qui est l'une des six langues officielles des Nations Unies et la langue maternelle de plus de 400 millions de locuteurs[1], ce qui représente environ 5.6 % de la population du monde entier. En juin 2017, le nombre d'utilisateurs arabes d'internet s'élevait à environ 185,000,000 millions, ce qui représente environ 43.8 % de la population du monde arabe et environ 4.8 % de la population du monde entier.

Cette quantité massive d'information dans la toile conduit à la nécessité de concevoir de nouvelles méthodes d'accès à ce volume croissant d'informations produites. Par conséquent, si elle doit être utilisée d'une manière plus économique et fructueuse, elle doit alors être organisée et facile à rechercher. 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].

1.2. Problématique

Les systèmes de recherche d'informations classiques reposent sur des approches qui représentent les documents (respectivement requêtes) par des descripteurs extraits à partir de leurs textes. Dans de tels systèmes, l'appariement document-requête est *lexical* basé sur la présence (respectivement absence) et la fréquence des mots de la requête dans le document. Or, les fréquences des mots clés dans un document et les entités nommées ne sont pas suffisantes pour identifier les relations exprimées et pour localiser les informations utiles pour une requête d'utilisateur en revenant aux documents initiaux (retour au contexte de

[1] https://www.internetworldstats.com/stats19.htm

l'information extraite). Alors, comment introduire « plus de sémantique » dans la recherche d'informations et dans la fouille textuelle ? Et quelle sémantique ?

Par ailleurs, la plupart des travaux de recherche dans le domaine de RI sémantique se sont orientés vers des documents textuels latins et peu d'études ont été effectuées sur des documents en arabe dont les caractéristiques grammaticales et morphologiques complexes rendent la tâche du traitement automatique plus difficile encore. En outre, la rareté et la faiblesse de ressources linguistiques moderne et de corpus d'expérimentations et d'évaluations en arabe posent plusieurs défis à l'avancement de la recherche dans ce domaine.

Surmonter ces limites est l'objet de plusieurs recherches récentes. C'est le cas notamment des approches de RI sémantique dans les documents textuels arabes. Les travaux menés dans le cadre de notre thèse s'inscrivent dans cet axe-là. Il s'agit, plus précisément, de trouver un modèle de représentation des documents et des requêtes en utilisant les connaissances extraites de ressources lexicales.

Un Système de recherche d'information sémantique « basé-connaissance » se focalise sur la désambiguïsation sémantique des mots souffrants de l'ambiguïté. Il existe différentes approches pour résoudre ce problème : d'une part les approches supervisées, nécessitant des corpus d'entraînement étiquetés manuellement et, d'autre part, des approches non-supervisées[Navi09].

Le problème avec les approches ou algorithmes supervisés est le fait d'obtenir de grandes quantités de textes annotés en sens est très coûteux. De plus, la qualité de la désambiguïsation de ces approches est restreinte par les exemples utilisés pour l'entrainement. C'est pourquoi les méthodes non supervisées sont intéressantes. Elles ne nécessitent pas de corpus annotés, et elles se divisent en deux catégories : d'une part les approches non supervisées qui exploitent les données non annotées ; et d'autre part les approches à base de connaissances qui utilisent des connaissances extraites de ressources lexicales. Dans cette étude, nous nous intéressons à ces dernières.

Il y a différents aspects à considérer dans le cadre des approches dite « basés-connaissances » : d'abord la question essentielle des ressources lexicales arabes qu'il est possible de les utiliser ? ensuite, la question de comment exploiter la ou les ressources lexicales pour désambiguïser ?

1.3. Contributions

Les travaux présentés dans cette thèse se situent dans le contexte précis de la RI sémantique dans les documents textuels en langue arabe. Plus précisément, nos contributions portent sur l'étude des approches d'indexation sémantique existantes et sur la proposition d'un cadre pratique pour l'amélioration de SRI sémantique [TaAl12] [TaSA18a] [TaSA18b] :

1. Conception et implémentation d'un système de recherche d'informations dédié à la langue arabe basé sur une méthode hybride en phase de lemmatisation combinant trois techniques connues : la suppression d'affixes proposée par Kadri[KaNi06a], les dictionnaires [AlEv94] et l'analyse morphologique [Bees98][Ahme00][MoMo02]. Ce système est qualifié d'être une base solide pour les systèmes de recherche d'information sémantiques arabes, due à sa méthode de lemmatisation puissante, car on ne peut pas désambiguïser, dans une phase ultérieure, un lemme incorrect.
2. Proposition d'un modèle d'indexation en ligne, qui nécessite une indexation semi-automatique. Ce modèle améliore les performances des systèmes d'extraction d'informations et de recherche d'information sémantique, car il aide ces systèmes à extraire et désambiguïser les mots composés.
3. Proposition d'une solution aux problèmes et carences dont souffre le traitement de la langue arabe, notamment en ce qui concerne la création de corpus, en développant un cadre d'application pour la création et le développement de corpus. En outre, notre thèse suggère une solution pour réduire les carences des systèmes d'évaluation de systèmes de la recherche d'informations, permettant aux chercheurs de tester leurs algorithmes d'indexation et de recherche.
4. Construction d'une nouvelle structure de ressource lexicale « DiLAC[2]» à partir de « dictionnaire de la langue arabe contemporaine »[مختا08] et montrer l'efficacité de cette ressource lexicale à la désambiguïsation en se basant sur la mesure de similarité sémantique.
5. Proposition d'un modèle de recherche d'information sémantique sur le texte arabe basé sur les connaissances, en utilisant « DiLAC » et l'algorithmes de Lesk simplifie.

[2]DiLAC : **Di**ctionnaire de la **L**angue **A**rabe **C**ontemporaine.

1.4. Organisation de la thèse

La thèse est organisée en six chapitres regroupés en deux parties. La première partie décrit et représente l'état de l'art sur la recherche d'information sémantique dans la langue arabe. La deuxième partie est destinée à la représentation et l'évaluation de nos contributions.

La première partie présente le contexte et les problématiques qui ont motivé nos travaux. Elle est composée de trois chapitres :

- Le premier chapitre présente les mécanismes de la RI, les concepts clés et les modèles de base sur lesquels repose la recherche d'information.
- Le second chapitre explique les SRI sémantiques et la désambiguïsation sémantique des mots (Word Sense Disambiguation : WSD) et présente les approches de WSD et leurs méthodes d'évaluations.
- Le troisième chapitre présente un état de l'art sur les différentes approches utilisées dans la recherche d'information sémantique dans les textes arabes et plus précisément le WSD et les caractéristiques de la ressource lexicale « le dictionnaire de la langue arabe contemporaine » que nous utilisons dans nos travaux.

La seconde partie présente nos contributions. Elle est composée de trois chapitres :

- Le quatrième chapitre présente l'architecture de notre système de recherche d'information dans un texte arabe, avant d'ajouter la couche sémantique, et fournit une analyse complète sur un certain nombre de niveaux liés à la recherche d'information, en particulier : (I) une étude des différentes méthodes de lemmatisation en arabe, (II) des applications de certaines méthodes de lemmatisation et (III) d'évaluation de la performance de notre contribution qui est une méthode de lemmatisation hybride pour la langue arabe.
- Le cinquième chapitre présente un nouveau type d'indexation pour contribuer à l'amélioration de la qualité des systèmes de RI arabes et pour construire des corpus de textes arabes appropriés pour mener les expériences nécessaires. La méthode d'indexation proposée appartient à la catégorie d'indexation semi-automatique et se compose de deux types. Le premier type effectue une indexation en ligne et le second type - sous cette méthode - est une indexation hors ligne (offline).

- Le sixième chapitre explique la méthode d'indexation sémantique implémentée pour la recherche d'information dans les documents textuels arabes où nous utilisons le DiLAC comme ressource pour explorer l'impact du passage d'une indexation basée sur des mots simples à une indexation basée sur des concepts. Ce chapitre décrit l'architecture de notre système, et présente l'expérimentation avec une discussion des résultats obtenus.

Finalement nous concluons notre travail par une conclusion générale et l'ouverture des perspectives essentielles qui sont susceptibles de l'enrichir d'avantage et d'affiner nos contributions.

Partie 1 :

État de l'art

Chapitre 1 :

La recherche d’information

1. LA RECHERCHE D'INFORMATION

1.1. Introduction

Parmi les caractéristiques inhérentes de l'ère de l'information est la croissance rapide et l'explosion de la quantité d'information hétérogène ; ce qui conduit à la nécessité de concevoir des approches pour enregistrer, filtrer, restituer et gérer cette quantité illimité d'informations.

Actuellement, la plupart des organisations ont un nombre important et croissant de documents, qui contiennent des informations d'une grande valeur potentielle. Ces types d'informations, d'un point de vue syntaxique et terminologique, sont structurées et bien organisées, en accord avec les directives politiques de l'organisation. Cependant, avec l'avènement d'Internet, les fournisseurs d'informations mettent à disposition un riche univers de connaissances sous différents formats, notamment le textuel. Ainsi, ce type d'information n'est ni structuré ni bien organisé. En fait, les informations disponibles se présentent sous différentes formes, on trouve des documents structurés, semi-structurés et/ou non structurés. Le domaine de la recherche d'information textuels « RIT » traite normalement la recherche de documents semi-structurés et non structurés. Les documents semi-structurés proviennent généralement de sources qui n'imposent pas une structure rigide (World Wide Web) ou lorsque les données sont combinées à partir de sources hétérogènes [AbVi97], [CHSA95], [LeRO96], [QRSU95].

Les utilisateurs, qui ont besoin d'informations, font face à un problème de surcharge d'informations, ce qui nécessite des technologies appropriées pour accéder à ce volume croissant d'informations produites. Cela a donné naissance au domaine de RI. La RI classique traite principalement des documents non structurés, qui consistent principalement en une forme libre de langage naturel, qui n'est pas toujours bien structuré et peut être sémantiquement ambiguë. Ainsi, les tendances actuelles dans ce domaine traitent de la recherche d'information sémantique telles que la recherche d'information à base d'ontologies et la recherche personnalisée d'informations basée sur le profil utilisateur.

La RI sur le Web présente des défis techniques supplémentaires par rapport à la RI classiques en raison de l'hétérogénéité et de la taille du Web. De plus, la RI sur le Web est exceptionnelle en raison des mises à jour continues et périodiques des informations, de la variété des langues utilisés, des duplications, des hyperliens, des requêtes mal formées et de la nature des utilisateurs.

Dans ce chapitre, nous allons d'abord commencer par présenter l'histoire de la RI, avant d'arriver à une définition adéquate de la RI et de déterminer les objectifs d'un SRI et ses composants. Ensuite nous allons donner une description des principaux modèles de RI. Cependant, il faut noter que nous restreignons cet état de l'art aux approches de RI les plus populaires. Enfin, ce chapitre termine par décrire les mesures d'évaluation communément utilisées dans le contexte de la RI.

1.2. Histoire de la RI

L'histoire de la RI remonte au 14ème siècle quand les premières bibliothèques avaient un certain nombre de livres, avec des informations qui devaient être restituées. A cet effet, les catalogues manuels ou les processus d'indexation ont été apparus. Avec l'avènement des ordinateurs au dix-neuvième siècle, ces catalogues étaient stockés comme données dans des bases de données. Quelques années plus tard, les SRIs ont été apparus, où pour la première fois le contenu informationnel du document était représenté par des mots-clés pour faciliter la tâche de leur restitution dans une phase postérieure.

[Luhn58] a décrit une technique simple, spécifique aux articles scientifiques, qui utilise la distribution des fréquences de mots dans le document pour pondérer les phrases et indiquer un degré de signification. Ces mots, appelés mots indexés, ont été utilisés pour représenter le document. L'idée de Luhn est d'utiliser des techniques statistiques pour la représentation des documents a eu un impact considérable, la grande majorité des systèmes de RI d'aujourd'hui étant basés sur ces mêmes idées.

À cette même période, le modèle booléen était plus accepté parce qu'il utilisait les expressions booléennes pour présenter les documents et les requêtes, mais il ne prenait pas en considération la pondération et le classement des mots-clés. Afin de résoudre ce problème, un schéma de pondération simple pour les "mots-clés" plutôt que les uns et les zéros a été conçu. [MaKu60] ont exploité, de nouveau, la méthode statistique afin de proposer une indexation probabiliste pour la recherche d'information. Le projet SMART [ClMK66] mettait en œuvre l'un des premiers systèmes l'indexation automatique.

En 1975, [SaWY75] a proposé un modèle vectoriel pour l'indexation automatique. Ce modèle a représenté les vecteurs de requêtes et des documents dans l'espace multidimensionnel, ainsi, il a utilisé le vecteur de requête et les mesures de similarité pour classifier des documents comme pertinents et non pertinents. Salton a également proposé un modèle booléen pondéré qui est une extension du modèle booléen standard, afin de généraliser le modèle vectoriel, qui

consiste principalement à pondérer les termes des documents au moyen d'un schéma tel celui du *Fréquence de Terme-Fréquence de Document Inverse* (TF-IDF de l'anglais Term Trequency-Inverse Document Frequency). Un autre système de pondération a été proposé par [Jone71], où ils ont montré expérimentalement qu'un terme est important dans un document donné, s'il apparaît souvent dans ce document et que peu de documents le contiennent.

Dans les années 1980, les chercheurs ont commencé à utiliser les techniques de traitement du langage naturel (Natural Language Processing ou NLP) pour extraire la sémantique de mots à partir d'un document plutôt que de simples chaînes de caractères isolées. Le NLP représente une tâche difficile, en particulier dans le SRI où un traitement doit être effectué pour représenter efficacement la sémantique du document. Les techniques d'indexation motivées linguistiquement (Linguistically Motivated Indexing ou LMI) [SLPW99] s'avèrent plus efficaces que les approches statistiques [MBSC97] lorsque des techniques de NLP plus avancées ont été proposées, telles que l'extraction de concepts, l'identification de phrases, la désambiguïsation des termes, la catégorisation, etc. Ainsi que, l'utilisation des techniques syntaxiques et sémantiques pour effectuer la « désambiguïsation du sens du mot » afin de trouver le bon sens du mot indexé et d'améliorer les performances RI.

Une autre méthode pour améliorer l'efficacité de la RI est la recherche basée sur le contenu, qui utilise le NLP pour extraire des connaissances à partir des documents. En 1995, dans CUM-6[3], les participants ont discuté la tâche qui forment la base de l'extraction d'information (Information Extraction ou IE). La définition de la pertinence est donnée implicitement par le modèle IE qui spécifie la connaissance lexicale spécifique au domaine, l'extraction, les règles et une ontologie (spécification formelle d'une compréhension partagée du domaine d'intérêt). Un système IE renvoie à l'information d'utilisateur répondant à la demande d'information de l'utilisateur où la demande est assez structurée et bien définie. Les informations obtenues à partir de IE peuvent être utilisées pour représenter le contenu du document et peuvent être utilisées à des fins d'indexation.

1.3. Qu'est-ce que la recherche d'information (RI) ?

La signification du terme « recherche d'information » est très vaste. En 1951, Calvin Mooers a inventé le terme « recherche d'information » pour décrire le processus par lequel un

[3] MUC-6, la sixième conférence sur la compréhension des messages, qui s'est tenue en novembre 1995. Cette conférence, qui a impliqué l'évaluation des systèmes d'extraction de l'information appliqués à une tâche commune, a été financées par l'ARPA pour mesurer et favoriser les progrès en matière d'extraction de l'information (Voir https://cs.nyu.edu/faculty/grishman/muc6.html).

utilisateur d'information potentiel peut convertir une demande d'information en une collection utile de références [Mooe50].

Selon Calvin Mooers : « La recherche d'information englobe les aspects intellectuels de la description de l'information et ses spécifications pour la recherche, ainsi que tous les systèmes, techniques ou machines qui sont utilisés pour effectuer l'opération. »

La RI vise à modéliser, concevoir et mettre en œuvre des systèmes capables de fournir un accès rapide et efficace, basé sur le contenu, à de grandes quantités d'informations [BaRi99]. Le but d'un SRI textuel est d'estimer la pertinence des documents textuels au besoin d'information d'un utilisateur. Un tel besoin d'information est représenté sous la forme d'une requête, qui correspond généralement à un sac de mots. Les utilisateurs ne sont intéressés que par les documents pertinents pour leurs besoins d'information. La représentation et l'organisation des informations doivent permettre à l'utilisateur d'accéder facilement aux informations qui l'intéressent.

Pour être plus efficace, le SRI doit en quelque sorte « interpréter » le contenu des documents d'un corpus et les classer selon un certain degré de pertinence par rapport à la requête de l'utilisateur. Cette « interprétation » du contenu du document implique l'extraction d'informations syntaxiques et sémantiques à partir du texte du document et l'appariement entre ces informations et les requêtes de l'utilisateur. La notion de pertinence est au centre de la RI [Trot05]. La partie la plus difficile du processus de recherche est de décider quels documents satisfont une certaine requête. Les documents devraient de préférence être classés par ordre décroissant de pertinence. Un SRI atteint son efficacité de recherche maximale lorsque les documents pertinents par rapport à la requête sont classés plus haut, tandis que les non pertinents sont classés plus bas.

1.4. Objectifs du SRI

L'objectif général d'un SRI est de minimiser la surcharge d'un utilisateur cherchant les informations nécessaires. Le temps système peut être exprimé comme le temps passé par un utilisateur à lire un élément contenant les informations nécessaires (par exemple, génération de requête, exécution de requête, analyse des résultats de requête pour sélectionner des éléments à lire, lecture d'éléments non pertinents). Le succès d'un système d'information est très subjectif, sur la base des informations nécessaires et de la volonté d'un utilisateur d'accepter les frais généraux. Dans certaines circonstances, les informations nécessaires peuvent être définies comme toutes informations contenues dans le système.

1.5. Composants du SRI

La recherche d'information est généralement un processus en deux étapes (Figure 1.1) :

- Les documents de corpus sont indexés.
- Puis les documents trouvés sont classés.

Le processus d'identification consiste à identifier les documents potentiellement pertinents à partir de l'ensemble de tous les documents. Ainsi, les documents pertinents sont ceux qui contiennent tous ou partie des éléments de recherche.

Le classement implique la combinaison d'un ensemble d'heuristiques dérivées du corpus, de l'ensemble de résultats et de documents individuels. Les heuristiques typiques incluent TF-IDF, mesures de proximité, etc. La similarité de chaque document à la requête est calculée et les documents sont triés en fonction de classement basé sur ces heuristiques.

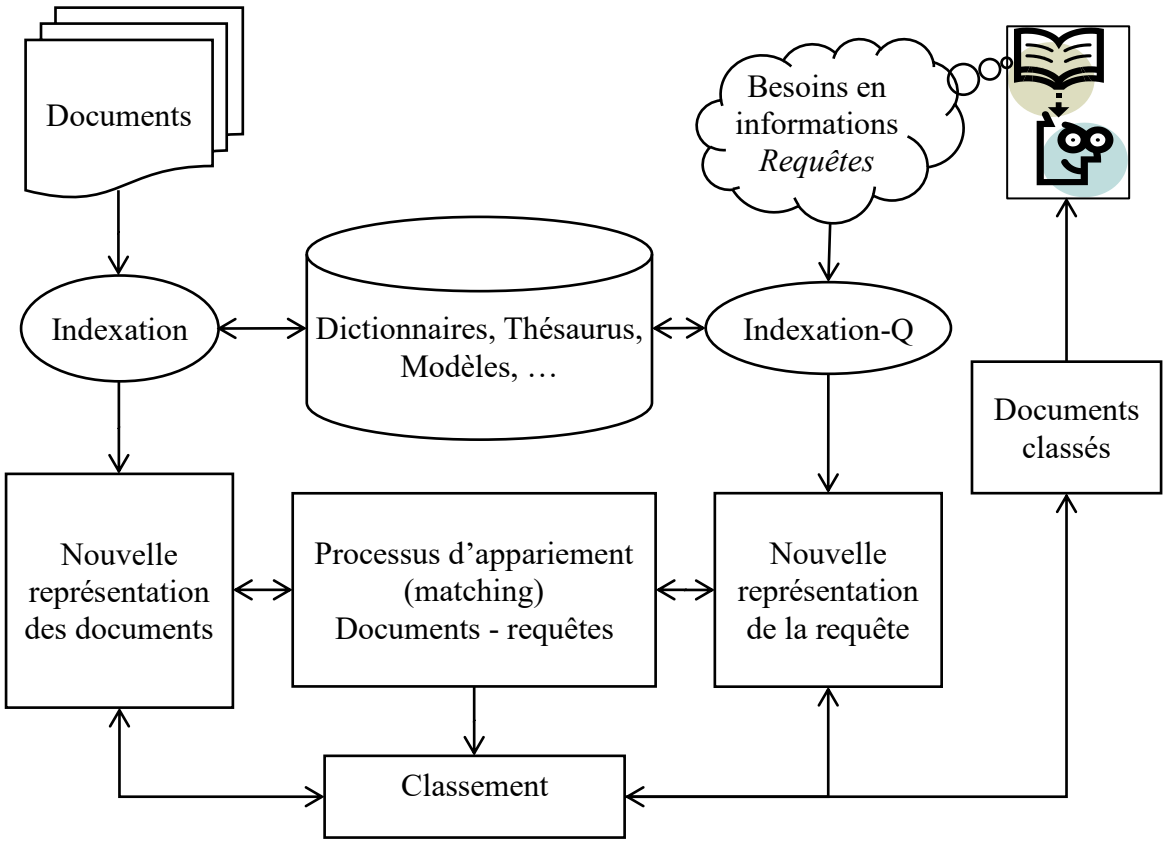

Figure 1.1: Architecture de SRI.

1.5.1. Indexation

Afin de juger efficacement si les documents d'un corpus correspondent à une requête donnée, on applique généralement un prétraitement appelé *indexation*. C'est la façon dont les documents sont gérés dans le corpus. Alors, un SRI enregistre les documents dans une représentation abstraite. Un ensemble de mots-clés est également enregistré, avec des liens vers le document dans lequel chaque mot apparaît. Cette structure d'enregistrement des informations d'indexation est appelée fichier inversé (Inverse File ou IF). Bien qu'il existe d'autres structures, IF demeure la structure de données la plus populaire utilisée par les SRIs. IF est une représentation de la collection de documents originale, organisée en listes appelées *Posting list*. Chaque entrée du fichier inversé contient des informations sur un seul terme de la collection de documents. Étant donné que cette structure nécessite une grande quantité d'espace à stocker, les *Posting list* sont généralement compressées.

Le processus d'indexation comprend plusieurs étapes, qui sont décrites comme suit :

1.5.1.1. Segmentation

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

1.5.1.1.1. Définition

La segmentation est une étape nécessaire et signifiante dans le traitement du langage naturel. La fonction d'un segmenteur est de découper un texte 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 important sujet dans le traitement de la langue naturelle car elle est étroitement liée « à l'analyse morphologique » [ChTa96], alors, la connaissance morphologique doit être incorporée au segmenteur. La segmentation est encore plus importante dans le traitement des langues riches et complexes morphologiquement comme l'Arabe, où un mot simple peut comporter un lemme et jusqu'à trois clitiques.

1.5.1.1.2. Les types de segmentation

Il existe plusieurs niveaux d'analyse du texte qui permettent de repérer les différents éléments constituant le texte et d'en définir les frontières. Il 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 (Figure 1.2).

- **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** qui cherche 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 appelé aussi *chunking*.

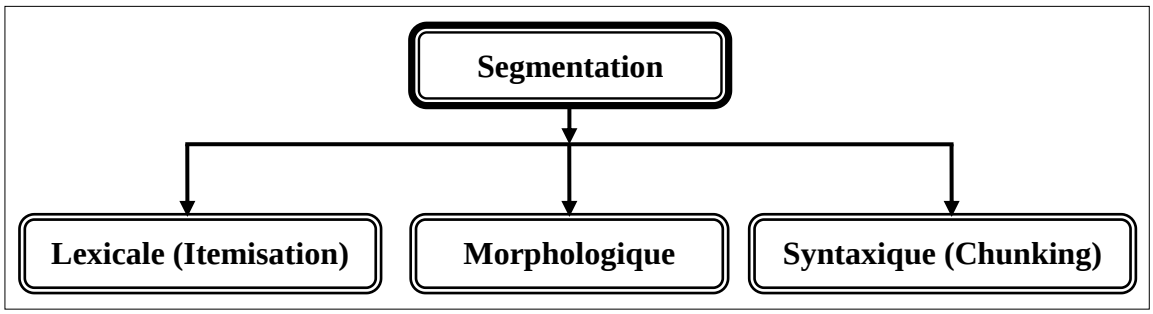

Figure 1.2: Les types de segmentation.

Parmi ces segmentations, nous étudions 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.

1.5.1.1.3. 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.

Les clitiques Arabes, cependant, ne sont pas reconnaissables facilement. Ils utilisent le même alphabet que celui des mots, sans la marque délimitant, et ils peuvent être enchaînés l'un après l'autre. Donc, sans connaissance morphologique suffisante, il est impossible de détecter et marquer les clitiques.

La segmentation Arabe est une étape préliminaire requise pour plusieurs applications de traitement automatique de langue naturelle. Elle a été décrite dans diverses recherches et a été mise en application dans beaucoup de solutions. Par exemple : l'analyse morphologique [Bees01], le diacritique [NeSh05], la recherche d'information [LaCo01a], et l'étiquetage de position [DiHJ04] [HaRa05].

1.5.1.2. Les mots vides

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

Luhn a souligné que la fréquence d'un mot dans un document peut être un bon discriminateur de sa signification dans le document [Luhn57]. En outre, il existe de nombreux mots extrêmement fréquents (par exemple "في" ou "Dans" en Français) qui apparaissent dans presque tous les documents d'un corpus. Ces mots, appelés mots vides (ou mots outils) et qui sembleraient être de peu de valeur pour représenter le contenu des documents, sont rejetés de la liste des termes d'indexation potentiels au cours du processus d'indexation [BaRi99]. La suppression des mots vides permet également de réduire la taille de l'index généré. Cependant, supprimer les mots vides d'un document à la fois prend du temps. Une approche rentable consiste à supprimer tous les mots qui apparaissent couramment dans le corpus et qui n'amélioreront pas le processus de la RI.

La stratégie générale pour déterminer une liste des mots vides est de trier les mots par sa fréquence dans le corpus (le nombre total d'apparition de chaque mot dans la collection de documents), et puis de prendre les mots 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 celle de mots vides alors ses termes sont ensuite rejetés lors de l'indexation.

Avec le temps, la tendance générale dans les systèmes de recherche d'information est l'utilisation standard des grandes listes des mots vides (200-300 mots). 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 RI s'est concentrée avec précision sur la façon dont nous pouvons exploiter les statistiques de la langue afin de mieux pouvoir faire face aux mots communs.

1.5.1.3. Normalisation

Certaines lettres du texte, représenté en langue naturelle, subissent une simple modification dans l'écriture qui n'influe pas sur le sens du mot. Mais l'encodage de ces lettres, dans la machine, change d'un mot à un autre. En outre, dans certaines langues, les voyelles sont souvent omises pendant l'écriture. Ce qui provoque la difficulté de pouvoir comparer ces mots. Afin de remédier à ce problème de variations du texte, il faut appliquer plusieurs genres de normalisation sur le texte et sur les requêtes.

Par exemple, dans l'arabe écrit, on peut utiliser deux lettres différents pour représenter un même mot. Dans l'arabe écrit aussi, 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 afin de pouvoir appairer les textes des requêtes avec ceux des documents du corpus (Tableau 1.1).

مكتبه	مكتَبة	مكُتبة

Tableau 1.1: Différentes écritures du mot « Librairie » en Arabe.

La normalisation des caractères est un processus qui peut améliorer le rappel. Un plus grand nombre de documents est extrait même si les documents ne correspondent pas exactement à la requête.

1.5.1.4. Lemmatisation et Racinisation

Souvent, un terme spécifié dans une requête d'un utilisateur peut prendre des formes variables dans les documents de corpus. Par exemple, en français la forme d'un verbe varie suivant le mode, le temps, la personne et le nombre. Par conséquent, ces différences, nécessaires pour certaines opérations comme l'étiquetage grammatical et l'analyse syntaxique, peuvent nuire à d'autres opérations. C'est le cas de la classification thématique de textes où il est préférable de traiter comme un lemme unique les différentes variantes issues d'une même forme canonique (Tableau 1.2).

Arabe	فكر	يفكر	نفكر	الأفكار
Français	pense	penser	pensons	Des idées

Tableau 1.2: Un exemple sur différentes variantes issues d'une même forme canonique.

De même, diverses ressources linguistiques comme les ontologies qui contiennent des règles dans lesquelles seules les formes canoniques sont présentes ; leur utilisation nécessite donc une étape préalable de mise sous forme canonique.

Deux opérations différentes peuvent être employées pour réduire cette variabilité de formes. La *lemmatisation* consiste à remplacer chaque mot (par exemple « الأفكار ») par sa forme canonique (« فكر »). La *racinisation* consiste à remplacer chaque mot (par ex. « الأفكار ») par sa racine (« فكار »). Notons que la racine n'est pas nécessairement un mot de la langue. La lemmatisation fait appel à l'analyse lexicale avec étiquetage grammatical. La racinisation emploie plutôt des règles simples de construction des mots dans une langue spécifique.

L'objectif de la lemmatisation/racinisation est de trouver la forme représentative d'index d'un mot à partir de sa forme représentée dans le document par l'application du processus de lemmatisation/racinisation. La question qui se pose pour ce processus est la suivante : Que doit-on choisir : lemmatisation ou racinisation ? Ou bien : quel lemme linguistique doit-on choisir à un mot pour une meilleure indexation ? Étant donné que, dans certaines situations, il n'est pas suffisant pour la RI de tronquer seulement un préfixe ou un suffixe de ce mot.

L'arabe est une langue fortement flexionnelle et a une structure morphologique complexe. La RI sur le texte arabe exige la forme de base du mot (racine ou lemme) pour être plus pertinente, donc le processus de lemmatisation est nécessaire (Tableau 1.3). 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).

<table>
<tr><td rowspan="4">(I) Avant la lemmatisation</td><td rowspan="2">La requête</td><td colspan="4">العرب</td></tr>
<tr><td colspan="4">les Arabes</td></tr>
<tr><td rowspan="2">Les documents</td><td>وعربي</td><td>كالعرب</td><td>العربي</td><td>والعرب</td></tr>
<tr><td>et Arabe</td><td>comme les Arabes</td><td>L'Arabe</td><td>et les Arabes</td></tr>
<tr><td rowspan="2">(II) Après la lemmatisation</td><td>La requête</td><td colspan="4">**عرب</td></tr>
<tr><td>Les documents</td><td>*عرب*</td><td>***عرب</td><td>**عربي</td><td>***عرب</td></tr>
</table>

Tableau 1.3: Un exemple sur le processus de lemmatisation d'un mot Arabe dans la RI.

Plusieurs lettres peuvent être attachées à un mot arabe, tandis qu'en français (en 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 RI.

Le processus de RI 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 RI, l'exemple précédent (Tableau 1.3) est utilisé.

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 SRI 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.

1.5.1.4.1. La structure de données d'index

Une structure de données appropriée est nécessaire, pour permettre un accès efficace aux documents de corpus. La structure de données la plus utilisée est l'index inversé, qui est un mécanisme orienté mot [BaRi99]. Le fichier inversé est un index lexicographique, c'est-à-dire une table alphabétique de mots-clés accompagnés de références. Il permet à partir d'un mot-clé donné de trouver toutes ses occurrences au sein d'un corpus. En général, et comme nous l'avons déjà mentionné dans la section 1 la structure d'index inversé comporte, pour chaque terme d'indexation, une liste appelée « *Posting list* » ou parfois « *Posting* » contenant l'identifiant des documents dans lesquels il apparait ainsi que sa fréquence d'apparition. Dans le cas où le fichier inversé mémorise en plus toutes les positions de chaque occurrence, le fichier inversé est dit : complet (*Full Inverted File*).

Cette structure permet de représenter, avec efficacité, l'ensemble de la collection des documents. En outre, elle diminue l'espace mémoire nécessaire et elle accélère la recherche, en conservant une seule occurrence de chacun des termes d'indexation.

$$< d >, < n >: \left[[< pos_1 > \#\{inf_1\}], [< pos_2 > \#\{inf_2\}], \dots, [< pos_n > \#\{inf_n\}]\right] \quad \text{(Eq.1.1)}$$

Où,

- $< d >$: nom du document ;
- $< n >$: fréquence du terme dans le document $< d >$;
- $< pos_{\text{i}} >$: i$^{\text{ème}}$ position du terme dans le document $< d >$;
- $\{inf_{\text{i}}\}$: autre informations sur le terme à la i$^{\text{ème}}$ position.

1.5.1.4.2. Analyseur de requêtes (Indexation-Q)

Cet analyseur exécute la segmentation, la suppression des mots vides et la lemmatisation de la requête afin de faciliter l'appariement entre le texte de la requête et les textes des documents du corpus.

1.5.1.4.3. L'appariement

Un SRI idéal devrait trouver les documents pertinents pour une requête donnée, et classer ces documents par ordre décroissant de pertinence. Dans cette phase, tous les documents contenant les termes de requête, dans sa version basique, sont extraits de la structure d'index inversée. La pertinence d'un document pour une requête donnée peut être évaluée par divers modèles de RI, tels que le modèle booléen, le modèle probabiliste et le modèle d'espace vectoriel (voir la section 1.6).

1.5.1.4.4. Classement

Enfin, tous les documents restitués sont classés en fonction de leur score de pertinence : le classement étant un composant primordial de la RI. Étant donné un ensemble d'objets ou d'instances, nous utilisons en général, un modèle de classement pour calculer le score de chaque objet et trier les objets en fonction des scores.

Dans la RI, les documents candidats doivent être classés en fonction de leur pertinence pour une requête. Cette tâche est pratiquement réalisée par une fonction de classement qui définit un ordre entre les documents en fonction de leur degré de pertinence par rapport à la requête de l'utilisateur. La fonction de classement est définie comme une fonction qui intègre les caractéristiques du document et attribue un score à chaque objet (en utilisant une fonction

de score f_{rang}), trie les objets par ordre décroissant des scores, et crée finalement le jeu de résultats $D^{[i]}$ contenant n documents pour la requête q_i.

Mathématiquement, le jeu de résultats $D^{[i]}$ est défini comme suit :

$$D^{[i]} = classer\left(f_{rang}(q_i, d_1), f_{rang}(q_i, d_2), \dots, f_{rang}(q_i, d_n)\right) \quad \text{(Eq.1.2)}$$

Plusieurs modèles de classement sont construits, par exemple, l'Okapi[4] BM25 (est une méthode de pondération utilisée en recherche d'information. Elle est une application du modèle probabiliste de pertinence, proposé en 1976 par Robertson et Jones[RoJo76]), le modèle booléen, le modèle d'espace vectoriel, le modèle probabiliste et le modèle de langage [Barw93], [LeTV93a], [LaGi98]. En plus de ces approches traditionnelles, les techniques d'apprentissage automatique, qui sont plus efficaces, sont de plus en plus utilisées pour le classement dans la recherche d'information.

1.6. Les modèles de la 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, la préparation de la scène pour résoudre les problèmes de la recherche d'information que nous allons essayer de traiter dans cette étude. En second lieu, fournir un aperçu rapide des principaux modèles de recherche d'information.

1.6.1. Les modèles classiques de la RI

Plusieurs modèles classiques de RI ont été développés et nous pouvons classer ces modèles comme suit :

1.6.1.1. Le modèle booléen

Le modèle booléen est l'un des plus anciens modèles de RI. Il est basé sur la théorie des ensembles et l'algèbre de Boole [Jone97]. Dans ce modèle, les documents sont représentés comme une conjonction logique de termes (non pondérés), c'est-à-dire, les termes sont combinés avec les opérateurs ET, OU et NON (Eq.1.3).

[4] Le terme « Okapi » faisant référence au nom du système de recherche de l'université de Londres où il a été implémenté initialement.

$$d = t_1 \wedge t_2 \wedge \ \ldots \ \wedge t_n \qquad \text{(Eq.1.3)}$$

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

Les requêtes sont également formulées comme expression booléenne des termes, c'est-à-dire, dans lesquels les termes sont combinés avec les opérateurs AND, OR et NOT.

ordinateur ET clavier ET (NOT souris)	$q = ordinateur \wedge calvier \wedge (\neg\ souris)$

Tableau 1.4: Un exemple sur les requêtes booléennes

Dans l'exemple ci-dessus, le modèle booléen extrairait tous les documents contenant à la fois les termes *ordinateur* et *clavier* mais pas le terme *souris*.

La relation de pertinence $R\ (d, q)$ entre une requête q et un document d est déterminée par les formules (Eq.1.4, Eq.1.5, Eq.1.6 et Eq.1.7) suivantes :

$$R(d, q_i) = \begin{cases} 1, & si\ q_i \in d;\ q_i\ est\ terme\ de\ q \\ 0, & si\ non \end{cases} \qquad \text{(Eq.1.4)}$$

$$\mathrm{R}(d, q_1 \wedge q_2) = \begin{cases} 1, & si\ R(d, q_1) = 1\ et\ R(d, q_2) = 1 \\ 0, & si\ non \end{cases} \qquad \text{(Eq.1.5)}$$

$$R(d, q_1 \vee q_2) = \begin{cases} 1, & si\ R(d, q_1) = 1\ ou\ R(d, q_2) = 1 \\ 0, & si\ non \end{cases} \qquad \text{(Eq.1.6)}$$

$$R(d, \neg q_1) = \begin{cases} 1, & si\ R(d, q_1) = 0 \\ 0, & si\ non \end{cases} \qquad \text{(Eq.1.7)}$$

1.6.1.1.1. Avantages du modèle booléen

Le modèle booléen a les avantages suivants :

- Il est facile à implémenter et son calcul est efficace [FrBa92].
- Il permet aux utilisateurs d'exprimer des contraintes pour décrire les caractéristiques linguistiques importantes [Marc91]. 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 [Coop88].

- 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.

1.6.1.1.2. Inconvénients du modèle booléen

Le modèle booléen standard souffre cependant des lacunes suivantes :

- Les utilisateurs éprouvent différentes difficultés pour construire des requêtes booléennes pertinentes pour plusieurs raisons [FoKo88]. 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. Ils sont aussi accablés par la multiplicité des moyens par lesquels une requête peut être structurée ou modifiée, en raison de l'explosion combinatoire des requêtes faisables quand le nombre de concepts augmente. Ils ont aussi du mal à identifier et appliquer les différentes stratégies qui sont disponibles pour rétrécir ou élargir une requête booléenne [LaWa93] [Marc91].
- Ce modèle ne retourne que 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'serait satisfait et le cas où tous (sauf un) seraient 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, trop de documents sont retrouvés souvent (le problème de « Output Overload »).

- Il est difficile de contrôler le nombre de documents recherchés : des utilisateurs sont souvent confrontés aux problèmes de (Null Output) ou (Output Overload) et ils sont perturbés de la façon de modifier la requête pour restituer 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 [Marc91].
- Le modèle ne représente pas le degré d'incertitude ou d'erreur due au problème de vocabulaire [BeCr92].

Tous ces inconvénients font du modèle booléen une option pour les SRI modernes, et plus précisément pour la recherche sur le Web.

1.6.1.2. Les modèles statistiques

Ces modèles utilisent des informations statistiques, comme la fréquence du terme et la fréquence du document inverse, pour déterminer la pertinence des documents par rapport à une requête. Ils existent deux modèles majeurs de la RI statistique sont :

1.6.1.2.1. 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, et 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 [BeCr92]. Le principe tient compte de l'existence d'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 et 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 (Eq.1.8) qui compare les deux probabilités :

$$O(D) = P(P/D)/P(NP/D) \qquad \text{(Eq.1.8)}$$

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 :

$$P(P/D) = P(D/P)\,P(P)/P(D) \qquad \text{(Eq.1.9)}$$

$$P(NP/D) = P(D/NP)\,P(NP)/P(D) \qquad \text{(Eq.1.10)}$$

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 :

$$\begin{aligned} O(D) &= P(P\ /\ D)/P(NP\ /\ D) \\ &= [P(D/P)\,P(P)/P(D)]/[P(D/NP)\,P(NP)/P(D)] \end{aligned} \qquad \text{(Eq.1.11)}$$

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

$$O(D) \simeq P(D\ /\ P)/P(D\ /\ NP) \qquad \text{(Eq.1.12)}$$

$O(D)$ 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 donc utiliser $P(D\ /\ P)/P(D\ /\ NP)$ à la place de $O(D)$. Ainsi, définissons $O(D)$ comme $P(D\ /\ P)/P(D\ /\ NP)$.

1.5.1.2.1.1. Avantages du modèle probabiliste

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.

1.5.1.2.1.2. Inconvénients du modèle probabiliste

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 [Spoe93].
- 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

[SaBu90]. Il y a plusieurs façons de calculer et de mettre à jour les poids et chacune a ses avantages et ses inconvénients.

1.6.1.2.2. Le Modèle vectoriel

Le modèle d'espace vectoriel représente un ensemble de documents en tant que vecteurs de poids dans un espace vectoriel (voir l'exemple de l'équation Eq.1.13), dont les dimensions sont les termes utilisés pour construire un index qui représente les documents [Dill83]. 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 [Dill83]. 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. Le poids de chaque terme peut être calculé en utilisant le schéma de pondération TF ou TF-IDF. Une requête peut également être vue comme un vecteur de document très court (voir l'exemple de l'équation Eq.1.14).

$$D = \begin{bmatrix} modèle & sémantique & recherche & SRI & NLP & \cdots & poids \\ 5 & 6 & 7 & 9 & 2 & \cdots & 1 \\ 1 & 1 & 0 & 0 & 5 & \cdots & 0 \\ 0 & 2 & 5 & 1 & 0 & \cdots & 2 \end{bmatrix} \quad \text{(Eq.1.13)}$$

$$R = [0 \quad 0 \quad 0 \quad 5 \quad 0 \quad \ldots \quad 1] \quad \text{(Eq.1.14)}$$

La similarité de cosinus entre le vecteur de requête et le vecteur de document pourrait être utilisée comme mesure du score du document pour cette requête. Par exemple, soit l'espace vectoriel suivant :

$$\langle t_1, t_2, t_3, \ldots, t_n \rangle \quad \text{(Eq.1.15)}$$

Un document (Eq.1.16) et une requête (Eq.1.17) peuvent être représentés comme suit :

$$d = \langle a_1, a_2, a_3, \ldots, a_n \rangle \quad \text{(Eq.1.16)}$$

$$q = \langle b_1, b_2, b_3, \ldots, b_n \rangle \quad \text{(Eq.1.17)}$$

Ainsi, a_i et b_i correspondent aux poids du terme t_i dans le document et dans la requête. Le degré d'appariement entre ces deux vecteurs est déterminé par leur *similarité* et il y a plusieurs façons de calculer la similarité entre deux vecteurs. En voici quelques-unes :

$$Sim_1(d,q) = \sum_i (a_i * b_i)\ (produit\ interne) \qquad \text{(Eq.1.18)}$$

$$Sim_2(d,q) = \sum_i (a_i * b_i) \Big/ \sqrt{\sum_i {a_i}^2 * \sum_i {b_i}^2}\ (cosinus) \qquad \text{(Eq.1.19)}$$

La deuxième (Eq.1.19) formule est normalisée, c'est-à-dire qu'elle donne une valeur dans [0, 1].

1.6.1.2.2.1. Avantages du modèle vectoriel

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.

1.6.1.2.2.2. Inconvénients du modèle vectoriel

Le 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.

1.6.1.3. Modèle logique

La RI doit inclure des formalismes capables de gérer l'incertitude et, en tant qu'outil permettant de saisir des connaissances imprécises et de raisonner sur ces connaissances, il n'y a probablement pas de meilleur formalisme que la logique. Les formalismes RI les plus prometteurs sont ceux qui combinent des approches bien connues pour évaluer l'incertitude, comme la théorie des probabilités, avec des paradigmes qui peuvent représenter des connaissances incertaines et permettre des inférences. Le principe d'incertitude logique de Rijsbergen sous-tend la plupart des modèles RI logiques.

1.6.1.4. Modèle linguistique

Les techniques RI visent principalement à détecter la pertinence, sans tenir compte des phénomènes linguistiques. Une technique RI est jugée efficace si elle peut différencier un texte

d'un autre. Cela a été fait assez bien en utilisant des méthodes quantitatives basées sur le nombre de mots et / ou de caractères, en particulier lorsque des distinctions de contenu relativement grossières étaient suffisantes. La technologie RI traditionnelle n'est pas motivée, pour la plupart, sur le plan linguistique. La croissance rapide de l'information dans le domaine de la technologie de l'information (NLP) a incité les gens à s'investir davantage dans les informations qu'ils fournissent. L'accent est mis sur le traitement linguistique des méthodes statistiques plutôt que sur leur remplacement. Du point de vue RI, le traitement linguistique est classé comme morphologique, lexical, syntaxique, sémantique et pragmatique. Une mesure minimale du traitement linguistique, en particulier aux niveaux morphologique et lexical, a été utilisée dans les techniques RI traditionnelles. Techniques de morphologie pour réduire les variantes d'un mot à une forme de racine commune, qui peut être considérée comme une méthode de normalisation. De même, les méthodes lexicales, y compris la construction d'une solution de bas niveau au problème, et l'utilisation de ces méthodes pour l'expansion du processus.

1.6.2. Modèles alternatifs

Ces modèles sont des améliorations des modèles classiques qui utilisent des techniques spécifiques à d'autres domaines. Le modèle booléen étendu[SaFW83], le modèle flou (Fuzzy) [Taha76], le modèle de réseau inférentiel [TuCr91], l'indexation sémantique latente (Latent Semantic Indexation : LSI) [DDFL90] sont des exemples de modèles de recherche d'information alternatifs (Figure 1.3).

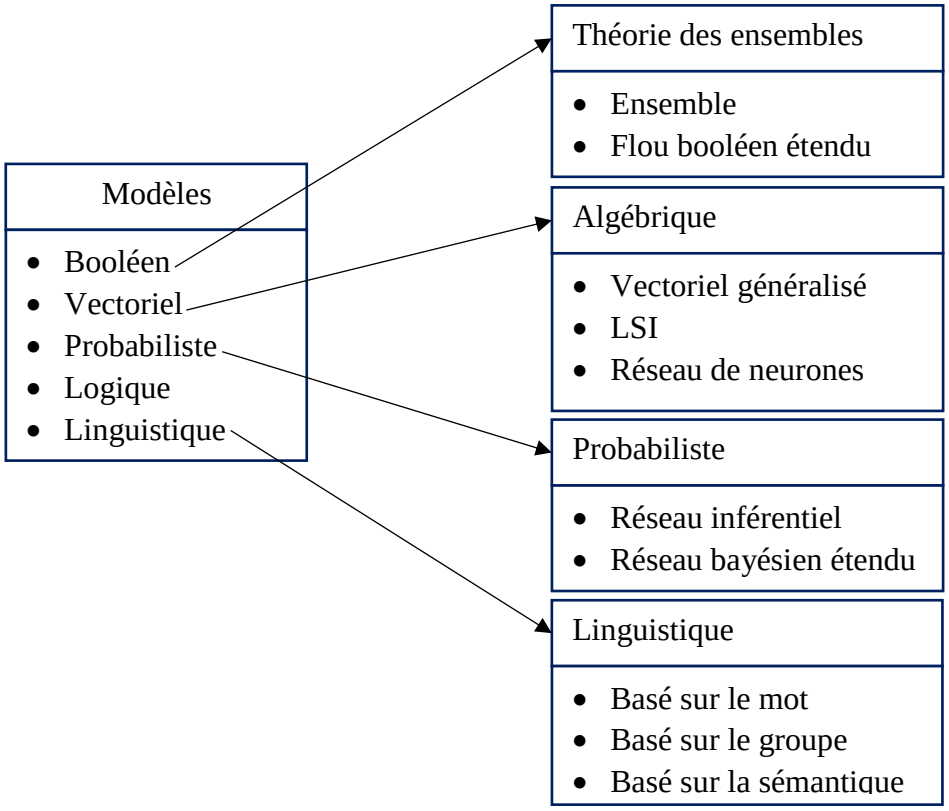

Figure 1.3: Modèles de RI alternatifs

1.7. Évaluation de SRI

Les mesures d'évaluation nécessitent une collection de documents (corpus) et un ensemble de requête. Toutes les mesures communes décrites supposent que chaque document de corpus doit être pertinent ou non pertinent pour une requête particulière [Barw93]. Les deux principales mesures généralement associées aux SRI sont la précision et le rappel. Lorsqu'un utilisateur décide de lancer une requête de recherche sur un sujet, la réponse sur cette requête est logiquement classifiée en quatre types illustrés à la figure (Figure 1.4). Les éléments pertinents sont les documents contenant des informations qui aident l'utilisateur à répondre à sa question. Les éléments non pertinents sont les éléments qui ne fournissent aucune information directement utile. Il existe deux possibilités pour chaque élément : il peut être restitué ou non par le SRI.

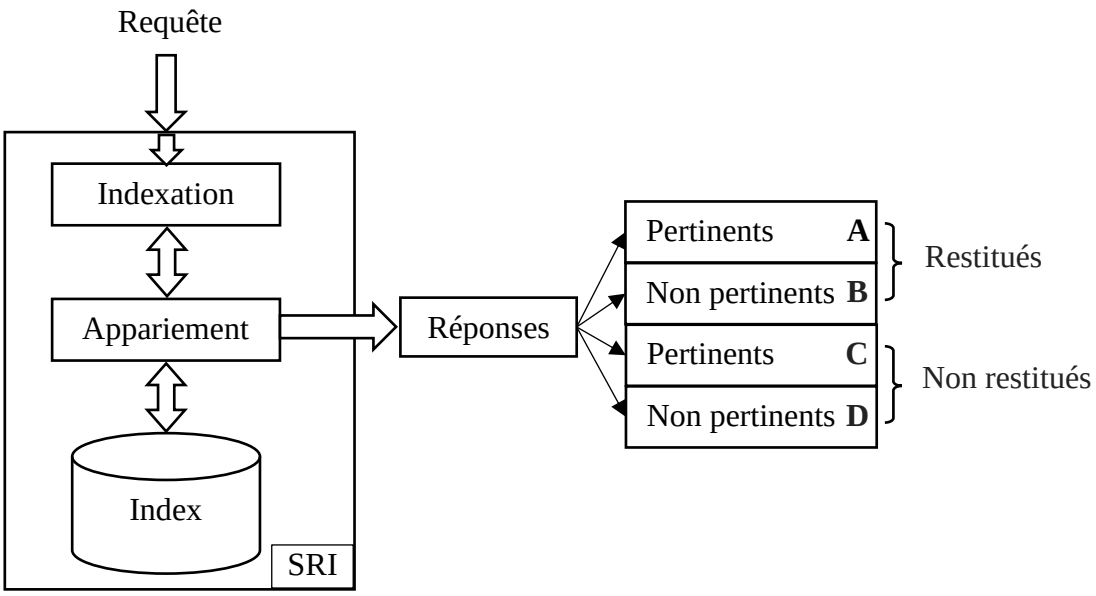

Figure 1.4: Classification des réponses de SRI aux requêtes des utilisateurs

1.7.1. Mesures de base

Les deux principales mesures utilisées pour évaluer un système de RI sont la précision et le rappel (Figure 1.5). Ces deux 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).

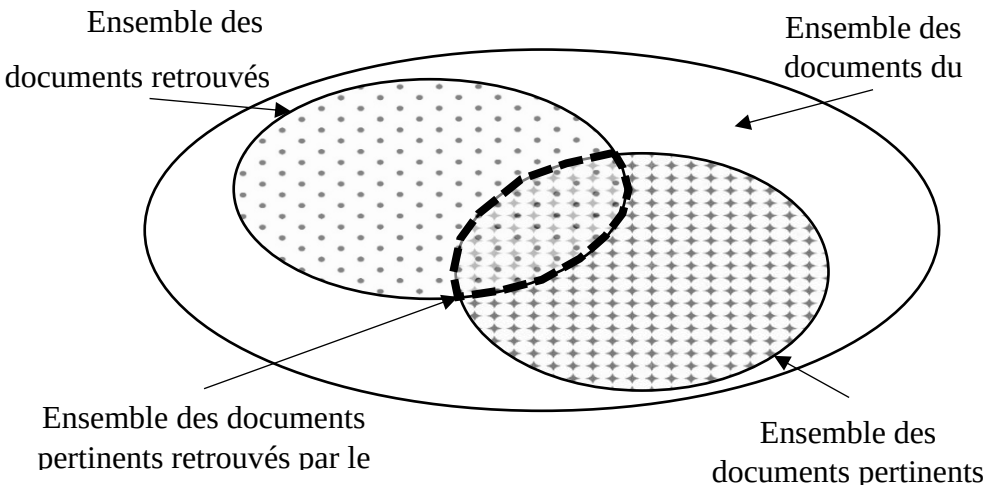

Figure 1.5: Précision et Rappel

1.7.1.1. 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 (Eq.1.20).

$$Précision = \frac{Nbre\ total\ de\ documents\ pertinents\ retrouvés\ par\ le\ système}{Nbre\ total\ de\ documents\ retrouvés\ par\ le\ système} \qquad \text{(Eq.1.20)}$$

En classification binaire, la précision est analogue à la valeur prédictive[5] positive[6]. La précision prend en compte tous les documents restitués. Elle peut également être évaluée à un seuil de coupure donné « N », en ne tenant compte que des résultats les plus élevés de « N » renvoyés par le système. Cette mesure est appelée précision à « N ».

La précision mesure un aspect de la surcharge de recherche d'informations pour un utilisateur associé à une recherche particulière. Si une recherche a une précision de 90%, 10% de l'effort de la part des utilisateurs consiste à passer en revue les éléments non pertinents.

1.7.1.2. Rappel

Un SRI 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.

$$Rappel = \frac{Nbre\ total\ de\ documents\ pertinents\ retrouvés\ par\ le\ système}{Nbre\ total\ de\ documents\ pertinents\ dans\ le\ corpus} \qquad \text{(Eq.1.21)}$$

En classification binaire, le rappel est appelé *sensibilité*. C'est le taux de vrais positifs, c'est à dire la proportion de positifs que l'on a correctement identifiés. C'est la capacité du modèle à détecter tous les incendies. Il peut donc être considéré comme la probabilité qu'un document pertinent soit restitué par la requête. Le rappel évalue dans quelle mesure un système traitant une requête particulière est capable de restituer les éléments pertinents que l'utilisateur souhaite voir. C'est est un concept très utile, mais en raison du dénominateur, il est impossible de le calculer dans les systèmes opérationnels. Si le système connaissait l'ensemble des éléments pertinents du corpus, il les aurait restitués.

[5] En statistique, ***la valeur prédictive*** d'un test est la probabilité qu'une condition soit présente en fonction du résultat de ce test. Le test doit être dichotomique, c'est-à-dire qu'il ne peut donner que deux résultats différents.
[6] ***La valeur prédictive positive*** est la probabilité que la condition soit présente lorsque le test est positif.

L'idéal pour un SRI est d'avoir de bons taux de précision et de rappel en même temps. Cependant, les deux métriques ne sont pas indépendantes.

1.7.1.3. La courbe de Précision/Rappel

Les performances d'un système de RI peuvent être représentées par une courbe Précision/Rappel. 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.

Il y a une forte relation entre les taux de précision et les taux de rappel : quand l'une augmente l'autre diminue. En pratique, la précision évolue en fonction du rappel et vice versa.

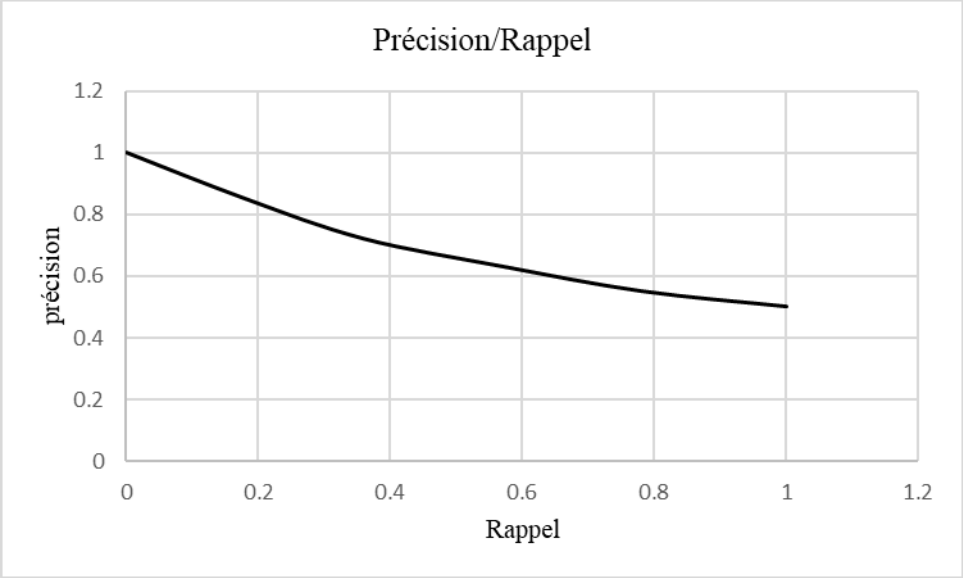

Figure 1.6: Courbe typique Précision/Rappel

1.7.2. Mesures alternatives

1.7.2.1. F- mesure (F-measure)

La F-mesure [Rijs79] 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 :

$$F_1 = 2\frac{Rappel(q).Précision(q)}{Rappel(q) + Précision(q)} \qquad \text{(Eq.1.22)}$$

Où :

q : est la requête.

Cette mesure F_1 donne la même importance à la précision et au rappel et des variantes (F_β) permettent de donner plus d'importance à l'un ou à l'autre.

$$F_\beta = (1+\beta)\frac{Rappel(q).Précision(q)}{\beta.Rappel(q)+Précision(q)} \quad \text{(Eq.1.23)}$$

Deux autres mesures F couramment utilisées sont la mesure $F_{0.5}$, qui pèse la précision deux fois plus que le rappel, et la mesure F_2, dont les poids rappellent deux fois la précision.

1.7.2.2. La précision moyenne (Mean average precision : MAP)

La précision moyenne (Ou la moyenne de la précision moyenne) est largement utilisée pour évaluer un système ou pour comparer des systèmes. Elle est définie par :

$$MAP = \frac{\sum_{r=1}^{N}(Précision(r) \times rel(r))}{Nombre\_Pertinente} \quad \text{(Eq.1.24)}$$

Avec $Nombre\_Pertinente$ le nombre de documents pertinents dans le corpus de documents, N le nombre de documents restitués, r le rang et $Précision(r)$ la précision lorsque les r premiers documents retrouvés sont considérés. $rel(r)$ vaut 1 si le document au rang r est pertinent et 0 sinon.

1.7.2.3. NDCG (Normalized Discounted Cumulative Gain)

NDCG est une mesure cumulative et multiniveau de qualité de classement. NDCG permet de considérer plus de deux niveaux de pertinence [JäKe02]. Pour une requête donnée, le NDCG est calculé comme suit :

$$N_i \equiv N_i \sum_{j=1}^{L} \frac{(2^{r(j)}-1)}{\log(1+j)} \quad \text{(Eq.1.25)}$$

Où $r(j)$ est le niveau de pertinence du $J^{\text{ème}}$ document et la constante de normalisation N_i est choisie de sorte qu'un ordre parfait résulterait en $N_i = 1$, L est le niveau de troncature de classement auquel le NDCG est calculé. Le N_i est alors moyenné sur l'ensemble de requêtes. NDCG est particulièrement bien adapté aux applications de recherche Web car il est à plusieurs niveaux et que le niveau de troncature peut être choisi pour refléter le nombre de documents affichés à l'utilisateur. Le gain cumulatif actualisé (DCG, Discounted Cumulative Gain) a été largement utilisé pour accéder à la pertinence dans le contexte de la recherche d'informations, car il peut gérer plusieurs niveaux de pertinence tels que {parfait, excellent, bon, passable}.

1.8. Conclusion

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

Le modèle joue un rôle central dans la RI. Il détermine le comportement clé d'un système de RI. Dans ce chapitre nous avons discuté les approches principales de modélisation de la recherche d'information. Aussi, 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.

Dans le chapitre suivant, nous compléterons notre présentation de RI du côté de modèle linguistique en nous penchant sur la recherche d'informations sémantique, en mettant le point sur les approches de désambiguïsation des sens des mots et leurs rôles dans la performance des systèmes de recherche d'informations sémantiques.

Chapitre 2 :

La recherche d’information sémantique et la désambiguïsation des sens des mots

2. LA RECHERCHE D'INFORMATION SEMANTIQUE ET LA DESAMBIGUÏSATION DES SENS DES MOTS

2.1. Introduction

Un texte est une série de mots perçus comme constituant un ensemble cohérent (phrase), porteur d'information (et connaissance) et utilisant les structures propres à une langue : conjugaisons, construction et association des phrases, etc. La représentation améliorée du texte faciliterait une restitution efficace d'information. Dans cette perspective, une représentation améliorée du texte devrait inclure non seulement des mots, mais aussi d'autres expressions, qui dénotent des entités, des concepts et des relations significatives [Strz99].

L'analyse syntaxique et sémantique sert à désambiguïser des phrases ambiguës, à extraire des connaissances, à interpréter des relations et des contenus. Elle permet aussi une meilleure représentation du texte. Dans ce chapitre, nous abordons l'effet d'introduire la sémantique dans le processus d'indexation sur l'amélioration de représentation de texte, par l'association de relations contextuelles entre les mots, qui sont utilisées à des fins d'inférence. L'intégration de la sémantique au processus de recherche d'information permet de passer d'une approche basée sur des mots, à la représentation des documents, à une approche basée sur la sémantique.

2.2. La sémantique et la recherche d'information

La recherche sémantique a pour objectif d'améliorer la précision de recherche par la compréhension approfondie des requêtes des utilisateurs, par la machine, et la signification exacte des termes tels qu'ils apparaissent dans le corpus, afin de générer des résultats plus pertinents.

La sémantique a été introduite à différents niveaux et à différentes étapes du processus de RI. Cependant, il convient de souligner que toute tentative d'apporter de la sémantique doit permettre d'équilibrer la quantité de traitement complexe du langage naturel nécessaire à l'augmentation des performances de RI.

Fondamentalement, un document peut être représenté avec un sac de mots (Bag of words) en utilisant le modèle booléen, qui ne fournit pas de classement des documents récupérés. Alors que l'on utilise l'approche statistique, le sac de mots est associé à des poids pour le classement [KoMa00] [BeBr05]. Le principal problème avec les approches traditionnelles de RI est que ceux-ci donnent de modestes résultats avec un taux de rappel et de précision moins élevé. Le modeste rappel est dû à la non-inclusion de synonymes de mots dans l'indexation, alors que la

modeste précision est due à la polysémie des mots. L'effet des aspects sémantiques tels que la polysémie (Un mot avec des significations différentes) et la synonymie (des mots différents ayant la même signification) sont les principales causes de la restitution des documents non pertinents. Le problème de la synonymie peut être résolu, d'une part, dans une large mesure grâce à l'utilisation d'une approche basée sur un synset[7] utilisant un thésaurus approprié. D'autre part, le problème de la polysémie peut être résolu à travers la révélation du sens correct d'un mot en utilisant un algorithme de désambiguïsation des mots (Word Sense Disambiguation : WSD) approprié. En terme général, WSD implique l'association d'une signification particulière (sens pouvant être potentiellement attribué au mot) à un mot dans un document, où les mots ont des significations différentes en fonction du contexte dans lequel ils apparaissent. L'utilisation des mots non ambigües aide grandement à améliorer les performances de RI.

En d'autres termes, les performances des RI peuvent être améliorées en incluant plus d'informations d'indexation sur les documents, telles que l'association de la signification sémantique avec les mots. L'idée de base est d'indexer la signification des mots plutôt que des mots simples. La sélection du sens le plus approprié pour un mot ambigu nécessite un travail supplémentaire fastidieux de traitement sémantique et la première tentative d'introduction de la sémantique se situait au niveau des mots, où [Voor93] [GVCC98] [MiMo00] tentait de dissocier les sens des mots.

L'introduction de la sémantique pour améliorer la désambiguïsation des mots a été considérée spécifiquement dans le contexte de la RI [Voor93] [GVCC98] [MiMo00]. Normalement, le sac de mots est d'abord déterminé et l'algorithme de désambiguïsation des mots est appliqué à ces mots afin de déterminer le sens exact de chaque mot. Ces sens sont utilisés avec le mot associé dans l'étape d'indexation. Les sens sont obtenus du thésaurus ou de toute ressource lexicale similaire. La raison pour laquelle WordNet a été choisi, dans la plupart des travaux de WSD, comme ressource lexicale par rapport à tout autre thésaurus en ligne est que WordNet fournit non seulement à l'utilisateur la signification d'un mot, mais fournit également des relations sémantiques telles que les synonymes, les antonymes, le hyponymes et les hyperonymes [Mill95]. La connaissance lexicale des mots en termes de ces relations permet de trouver le sens correct des mots indexés par le module WSD. Par conséquence, les

[7] Un ensemble de synonymes qui sont interchangeables dans un contexte donné sans changer la valeur de vérité de la proposition dans laquelle ils sont incorporés.

algorithmes de WSD basé sur les sens (Sense-based WSD) utilisent les sens lexicaux à des fins de désambiguïsation.

2.3. La désambiguïsation des sens des mots (Word Sense Disambiguation : WSD)

La désambiguïsation des mots joue un rôle important dans presque tous les domaines du traitement du langage naturel : traduction automatique, recherche d'informations, analyse des sens et reconnaissance vocale, compréhension des messages, communication homme-machine, navigation hypertexte, analyse thématique, analyse grammaticale, traitement de la parole, traitement de texte et système de réponse aux questions. C'est pourquoi le domaine de recherche WSD a une grande importance théorique et pratique.

Dans le traitement automatique des langues naturelles (Natural Language Processing : NLP), la désambiguïsation du sens des mots, communément appelée WSD, est l'identification du sens voulu d'un mot, compte tenu de son utilisation dans une expression linguistique plus large, c'est-à-dire une phrase. Dans la littérature de la lexicographie et la linguistique, les sens des mots sont connus comme des entités très glissantes.

Définition : La désambiguïsation des mots est la détermination du sens exacte d'un mot, dans une phrase donnée, ayant un certain nombre de sens distincts. Par exemple, pour un humain, il est évident que la phrase « *Cet ours a mangé un avocat* » utilise le mot « avocat » au sens de « fruit » et non pas au sens de « plaideur ». Cependant, le développement d'algorithmes pour reproduire cette capacité humaine est une tâche difficile. Ainsi, le problème majeur de la désambiguïsation des mots est la prise de décision concernant le sens approprié parmi d'autres. Particulièrement, dans les cas où les différents sens peuvent être étroitement liés (l'un étant une extension métaphorique ou métonymique d'un autre), et la division des mots en sens devient beaucoup plus difficile.

Une corrélation étroite a été trouvée entre la signification lexicale et sa distribution. Selon une étude dans le domaine des sciences cognitives [Lenc08], les personnes n'ont souvent besoin que de quelques mots supplémentaires dans un contexte donné (généralement un seul mot est suffisant) pour désambiguïser le sens des mots.

La relation entre un mot donné et d'autres mots dans son contexte peut être utilisée efficacement pour la désambiguïsation : le contexte fournit généralement les informations nécessaires pour ce processus. La première étape de la tâche de développement du modèle WSD consiste à déterminer le type d'informations contextuelles utiles et la taille du contexte requise.

Un cadre d'analyse syntagmatique à plusieurs niveaux peut être conçu pour décrire les contraintes syntaxiques et sémantiques du mot donné. Au cours de l'étude de 5793 mots, [TYBM00] ont constaté que différents sens ont des distributions différentes et complémentaires aux niveaux de la syntaxe et / ou de la collocation. Cela sert de base pour établir un modèle de désambiguïsation des mots en utilisant des informations grammaticales et un thésaurus fourni par les linguistes.

En résumé, le contexte approprié est l'outil de base pour travailler sur la désambiguïsation des mots, indépendamment du fait que cela soit réaliser par des humains ou par des machines.

La désambiguïsation automatique des sens des mots a suscité un intérêt et une préoccupation depuis les premiers jours du traitement automatique du langage naturel dans les années cinquante, la désambiguïsation des sens est considérée comme une « tâche intermédiaire », car il s'agit d'une tâche très nécessaire pour obtenir de bons résultats de toute tâche liée au traitement du langage naturel.

Il est instructif aussi de comparer le problème de désambiguïsation des mots avec le problème de l'étiquetage en parties du discours (Part of Speech Tagging : POS Tagging)[8]. Les algorithmes utilisés pour ce dernier ne tendent pas à bien fonctionner pour l'autre, parce que la POS d'un mot est principalement déterminée par les mots (de 1 à 3) immédiatement adjacents, tandis que le sens d'un mot peut être déterminé de manière un peu plus loin. Le taux de réussite des algorithmes de l'étiquetage en parties du discours est actuellement beaucoup plus élevé que celui de la désambiguïsation des mots, la précision étant d'environ 95% ou mieux, contre moins de 75% de précision avec l'approche d'apprentissage supervisé [BaVC06](voir Section 2.6.2). Ces chiffres sont typiques de l'anglais et peuvent être très différents de ceux des autres langues.

Le problème de la désambiguïsation du sens des mots a été décrit comme un problème de l'intelligence artificielle, c'est-à-dire un problème qui ne peut être résolu qu'en résolvant d'abord les problèmes difficiles de l'intelligence artificielle, tels que la représentation du sens commun et les connaissances encyclopédiques. Cela facilite la révélation du sens exact du mot dans son contexte.

Le rôle du module WSD est de suggérer les sens des mots ambigus et cela nécessite donc l'étude de la signification de différentes entités linguistiques telles que les mots, les phrases, les textes, etc. L'approche componentielle (également appelée dé-compositionnelle) définissent

[8] En linguistique, l'étiquetage en parties du discours consiste à attacher automatiquement une étiquette d'une partie de discours à tous les mots d'un corpus donné.

les unités lexicales en décomposant leur sens en unités sémantiques plus petites (des atomes) ou des primitifs sémantiques. Traditionnellement, la sémantique incluait l'étude du sens connotatif et de la référence dénotative, des conditions de vérité, de la structure des arguments, des rôles thématiques, de l'analyse du discours et du lien de tous ces éléments avec la syntaxe.

L'étude du sens des mots englobe l'étude des relations entre les mots tels que l'homonymie, la synonymie, l'antonymie, la polysémie, la paronymie, l'hyperonymie, l'hyponymie, la méronymie, la métonymie, et l'holonymie, et l'étude des relations exocentriques et endocentriques entre différentes expressions linguistiques. En outre, il s'agit de l'étude des rôles thématiques, de la structure des arguments et du lien avec la syntaxe, le traitement du sens et de la référence, des conditions de vérité et de l'analyse du discours.

Le développement des modèles de désambiguïsation des mots plus efficaces ne cesse d'évoluer d'une année à une autre en considérant de tels aspects théoriques de la signification des entités linguistiques, en particulier les mots.

2.4. Contexte historique

En remontant à la fin des années 1940 [Weav49], WSD a été conçu comme la tâche fondamentale de la traduction automatique (Machine Translation : MT), dans les années 1960, [Barh60] a également reconnu que le WSD est le principal obstacle au développement du système de MT. Dans les années 1970, [Wilk75] a étudié l'impact de la sémantique préférentielle sur la compréhension du langage naturel, qui spécifie des restrictions de sélection concernant les combinaisons d'éléments lexicaux dans une phrase, à l'aide de traits sémantiques. Au cours de cette période, il était difficile d'obtenir l'approche généralisée en raison du manque de grande quantité de connaissances lisibles par machine. Dans les années 1980, la recherche sur le WSD a connu une remarquable révolution à mesure que des ressources lexicales à grande échelle et des corpus devenaient disponibles. En conséquence, les chercheurs ont commencé à utiliser différentes procédures d'extraction automatique des connaissances [Wilk93] parallèlement aux méthodologies manuelle. L'application des approches statistiques pour établir une évaluation constante des systèmes de WSD était le principal objectif des années 1990 [WiSt96]. L'ère des années 2000 a marqué un progrès considérable dans différents domaines de recherche, en particulier dans le domaine des sources de connaissances et de leurs interactions [StWi01]. Par la suite, l'évaluation de divers algorithmes d'apprentissage, pour le WSD, a été réalisée [EdCo01] [WiFr00].

2.5. Informations nécessaires pour WSD

2.5.1. Le contexte

Le rôle du contexte dans tous les travaux de désambiguïsation est majeur. [IdVé98] a constaté que la seule manière d'identifier le sens d'un mot ambigu est de se référer à son contexte. Plus précisément, la sélection du sens exact d'une instance d'un mot ambigu consiste à éliminer les significations qui génèrent une « ambiguïté » sémantique avec le contexte de l'instance. A la fin de ce processus, si toutes les significations sont supprimées sauf une, celle-ci est sélectionnée et attribuée à l'instance en question. Ce mécanisme de désambiguïsation est approprié des méthodes automatiques de WSD.

Le contexte joue donc un rôle de *réducteur,* d'une part, parce qu'il opère une diminution de nombre d'ambiguïtés virtuelles. Ce rôle du contexte dépend des mots qui y apparaissent et l'inclusion dans le contexte de mots sémantiquement apparentés à l'un des sens du mot ambigu (également appelés *mots amorces*) aide à la sélection de ce sens parmi l'ensemble des sens possibles du mot. D'une autre part, le contexte peut jouer un rôle d'*inducteur*, dans le cas où il existe une parenté (affinité) préférentielle entre lui et l'une des significations de l'expression ambiguë, affinité que l'on peut traduire en termes de probabilité relative d'apparition de la signification en présence du contexte considéré [Fuch96].

2.5.1.1. Exploitation des informations du domaine

2.5.1.1.1. La désambiguïsation par restriction à des domaines précis

Les informations d'un domaine donné (tels que MEDICINE, ARCHITECTURE et SPORT) constituent un moyen naturel et puissant d'établir des relations sémantiques entre les sens des mots, qui peuvent être utilisés de manière rentable pendant le processus de désambiguïsation. En particulier, les informations d'un domaine spécifique sont fondamentalement caractérisées par la cohérence du texte, de sorte que les sens des mots apparaissant dans une partie du texte ont d'autant plus de chances d'être sémantiquement similaires. L'importance de l'information de domaine dans la WSD a été soulignée dans plusieurs travaux, notamment [GVCC98] [BuSa01].

L'impact des informations du domaine et du sujet traité dans la désambiguïsation est en effet important pour sélectionner le sens correct des mots. [Oswa52] [Reif54] ont proposé une solution pour la réduction des sens possibles des mots ambigus consistant à construire des microglossaires, qui sont des glossaires destinés à utiliser au sein des domaines spécialisés,

dont les sens des mots ambigus étaient réduits aux sens pertinents dans le domaine concerné, ce qui éliminait une partie de son ambiguïté.

Les informations d'un domaine donné peuvent être repérées, si elles ne sont pas fournies au sein de ressources spécialisées, à l'échelle du document ou dans des portions de texte plus petites. Pour chaque domaine, il existe un sous-vocabulaire de termes appropriés le désignant. Les méthodes de désambiguïsation qui utilisent ce type d'informations contextuelles exploitent la redondance dans les textes. Dans ces cas, le contexte est traité comme un sac de mots ; ce qui importe étant la cooccurrence d'un sens précis du mot ambigu avec des mots liés au sujet traité au sein d'une fenêtre textuelle. La tâche de désambiguïsation consiste alors à identifier le sujet traité par le texte et à sélectionner le sens adéquat du mot ambigu.

Plusieurs études ont été basées sur le domaine traité, comme la principale source d'informations pour la désambiguïsation. [GaCY92a] a expérimenté une approche de recherche d'information pour lever l'ambiguïté des sens. Ainsi, les contextes sont définis de manière à utiliser les 50 mots à gauche et les 50 mots à droite du mot polysémique en question. La désambiguïsation se base sur le principe « un sens par discours », ce principe régissant également la méthode proposée par [Yaro95a], selon lequel, le sens d'un mot est le même tout au long d'un document.

[MSPG01] ont décrit une approche de WSD basée sur des informations de domaine. L'hypothèse sous-jacente est que les domaines constituent un élément fondamental de la cohérence du texte. En conséquence, les sens des mots apparaissant dans une partie cohérente du texte ont tendance à maximiser la similarité de domaine. Trois systèmes ont été mis en œuvre, intégrant des techniques basées sur la connaissance et les statistiques. Concernant les ressources lexicales, les systèmes utilisent « WordNet Domains », qui est une extension de la version anglaise Wordnet 1.6, où les synsets ont été annotés avec des informations de domaine [MaCa00]. L'algorithme de désambiguïsation est basé sur la comparaison entre les vecteurs de domaine qui collectent des informations contextuelles relatives au mot cible et les vecteurs construits à partir de WordNet déterminant les domaines pertinents pour les sens des mots ambigus. Le sens correspondant au vecteur le plus proche au vecteur du contexte est alors sélectionné comme étant le sens approprié du mot.

2.5.1.1.2. Limites de l'apport du domaine pour la désambiguïsation

Portant que la méthode basée sur les informations du domaine (ou du sujet traité) peut aider à la désambiguïsation de manière importante, surtout dans le cas de sens lexicaux bien

distincts, elle rencontre une difficulté de traiter certains cas d'ambiguïté où les informations ne sont pas suffisantes pour sélectionner des sens plus apparentés [Jone64] [Yaro92] [LeCh98], étant donné le degré variable de spécialisation des textes et la diversité possible de sujets dans le même texte, et ce, même au niveau de petites paragraphe. En outre, les sens différents d'un mot à *polysémie logique*[9] [Pust91] peuvent paraître tous équivalents pour l'interprétation du mot dans un domaine, bien qu'un seul sens soit visé dans un contexte particulier.

[Jone64] a considéré les étiquettes de domaine comme non pertinentes pour la description du sens d'un mot dans un contexte particulier : un mot peut en effet être utilisé dans des sens différents dans un texte qui traite pourtant d'un sujet bien précis. [Yaro92] a souligné également les limites d'une telle méthode de désambiguïsation. Ces limites caractérisent, d'une part, les mots qui présentent des distinctions sémantiques indépendantes d'un sujet précis et, d'autre part, les cas où des distinctions sémantiques fines peuvent être repérées au sein d'une catégorie du thésaurus.

L'efficacité des méthodes de désambiguïsation basées sur les informations de domaine est alors influencée par la sémantique des mots traités. Des divergences quant à l'efficacité de ces méthodes peuvent aussi être observées au niveau d'un seul mot ; ainsi lorsque le mot est caractérisé conjointement par homonymie et polysémie.

2.5.1.2. Le contexte local ou « micro-contexte »

L'inadéquation constatée des informations de domaine à la désambiguïsation dans un grand nombre de cas a généré la recherche d'autres sources d'informations plus appropriées, dont la plus importante est le contexte lexical ou local des mots. La plupart des travaux de désambiguïsation utilisent le contexte local d'une occurrence de mot comme source d'information principale pour WSD.

Définition 1 : Le contexte local ou « micro-contexte » est généralement considéré comme une petite fenêtre de mots entourant le mot cible dans un texte ou un discours, de quelques mots jusqu'à la phrase entière dans laquelle le mot cible apparaît.

Définition 2 : Le contexte est très souvent considéré comme l'ensemble des mots ou des caractères apparaissant dans une fenêtre du mot cible, sans égard pour la distance, la structure syntaxique, ou d'autres relations.

[9] Pustejovsky a définit la polysémie logique comme une ambiguïté complémentaire dans laquelle il n'y a pas de changement de catégorie lexicale et où les sens multiples du mot ont des significations qui se chevauchent, dépendantes ou partagées.

Le contexte local est généralement délimité par une fenêtre textuelle qui se situe à gauche ou à droite ou des deux côtés du mot cible et dont la taille peut varier. La définition de la taille de la fenêtre textuelle dépend de celle de la distance optimale entre le mot cible et les indices contextuels pouvant servir à sa désambiguïsation.

2.5.1.2.1. Taille du contexte

Selon [Weav49], il est impossible de déterminer le sens d'un mot dans un livre, si l'on analyse ce mot séparément des autres mots, comme à travers un masque opaque avec une fente de la taille d'un mot. Cependant, si on élargit la fente du masque, jusqu'à ce que l'on puisse voir N mots de chaque côté du mot cible, alors si N est assez grand, on peut connaître avec certitude du sens du mot cible.

La détermination de la distance optimale a fait l'objet d'un grand nombre de travaux dont les résultats sont assez variés.

Pour [Kapl55], par exemple, le mot précédant le mot polysémique dans le texte est un très mauvais indice de désambiguïsation et moins approprié que le mot suivant. Une fenêtre comprenant un mot de chaque côté du mot polysémique (c.à.d. N = ± 1) est plus efficace que celle qui en contient deux (c.à.d. N = ± 2), et l'intérêt de retenir deux mots de chaque côté du mot polysémique est comparable à celui de la phrase entière. [ChLu85] ont confirmé les conclusions de Kaplan selon lesquelles les taille de la fenêtre N = ± 1 ou N = ± 2 mots autour du mot cible sont très fiables pour la désambiguïsation, mais essentiellement pour la désambiguïsation des homographes. Cependant, malgré ces résultats, la valeur de N a continué à varier de façon plus ou moins arbitraire au cours des travaux de WSD.

Yarowsky [Yaro93] [Yaro94] [Yaro99] a examiné différentes fenêtres de micro-contexte (N = ± k) et les trie en utilisant un ratio *log- likelihood* (log-vraisemblance) pour trouver la taille la plus optimale pour la désambiguïsation. Yarowsky a observé que la valeur optimale de k varie selon le type d'ambiguïté, il suggère que les ambiguïtés syntaxiques locales n'ont besoin que d'une fenêtre de N = ± 3 ou N = ± 4, tandis que les ambiguïtés sémantiques ou thématiques nécessitent une fenêtre plus large (de 20 à 50 mots).

En outre, Yarowsky (*ibid.*) a également utilisé d'autres informations (telles que la catégorie grammaticale) qui constituent un autre facteur de variation de la taille de la fenêtre : une grande fenêtre textuelle peut être utilisée pour la désambiguïsation des noms, mais pour les verbes et les adjectifs, sa taille doit être beaucoup plus petite. En effet, [GaCY92a] [GaCY92b] montrent que l'utilisation d'un contexte large (±50 mots autour du mot

polysémique) améliore sensiblement les résultats de la désambiguïsation des noms polysémiques, par rapport à l'utilisation d'un contexte plus restreint (±6 mots).

[LeMC98] utilisent une fenêtre de ± 3 mots autour du mot ambigu, tandis que le classifieur utilisé par [BrWi94a] [BrWi94b] prend en compte ± 2 mots autour du mot ambigu.

2.5.1.2.2. La collocation

La collocation constitue un autre paramètre qui a été utilisé de différentes manières dans les travaux WSD. Le terme a été popularisé par J. R. Firth dans son article de 1951 intitulé « Modes de signification » [Firt75]. Il souligne que la collocation n'est pas une simple cooccurrence mais est une « habituelle cooccurrence ». [Hall61] pense que les collocations sont des combinaisons entre des unités lexicales qui dépassent la limite de la grammaire et accorde un rôle central aux collocations dans l'apprentissage du lexique.

Pour Halliday (*ibid.*), l'analyse sur le plan lexical des textes, se base essentiellement sur les collocations. Le texte est vu non comme un ensemble d'unités lexicales, mais comme une unité sémantique à part entière. La collocation possède une fonction de cohésion lexicale, et a pour impact la cohésion textuelle, c'est-à-dire la propriété qu'a le texte de ne pas être agencé par des phrases quelconques.

Halliday (*ibid.*) a défini la collocation comme : « L'association syntagmatique des éléments lexicaux, quantifiables, textuellement, comme la probabilité qu'il y aura à n suppression (sur une distance de n éléments lexicaux) à partir d'un élément x, les items a, b, c [...] ».

[Berr73] a basé sur les la définition de Halliday et a proposé cette définition : « une collocation significative peut être définie comme une association syntagmatique entre des éléments lexicaux, dans laquelle la probabilité de cooccurrence de l'élément x avec les éléments a, b, c ... est supérieure au hasard ». C'est dans ce sens que la plupart des travaux WSD utilisent le terme.

Le rôle important des « mots amorces » sur la sélection du sens correct d'une instance d'un mot polysémique a été démontré par des expériences menées en psycholinguistique, qui mesurent la facilité avec laquelle le sujet sélectionne le sens correct du mot en présence de mots amorces situés à proximité [KiMr85]. Les résultats de ces expériences ont montré que les collocations sont traitées de façon différente des autres cooccurrences : les mots amorces qui se trouvent en collocation fréquente avec les mots polysémiques servent à les activer dans les

tâches de décision lexicale, tandis que ceux qui sont liés au contexte thématique ne facilitent pas les décisions lexicales des sujets.

[Yaro93] aborde explicitement l'utilisation de collocations dans le travail WSD, mais il convient d'adapter la définition à son objectif comme « la cooccurrence de deux mots dans une relation définie ». Il examine une variété de relations de distance, mais considère également la contiguïté par POS (par exemple, premier nom à gauche). Il a déterminé que dans les cas d'ambiguïté binaire, il existe « un sens par collocation », c'est-à-dire que, dans une collocation donnée, un mot est utilisé avec un seul sens, avec une probabilité de 90 à 99%.

2.5.1.2.3. Les relations syntaxiques

Les relations syntaxiques constituent un autre facteur de désambiguïsation. [Earl73] a utilisé la syntaxe exclusivement pour la désambiguïsation dans la traduction automatique. A ce jour, la plupart des travaux WSD ont utilisé les informations syntaxiques conjointement avec d'autres informations. L'utilisation de restrictions de sélection pèse lourdement dans les travaux basés sur l'intelligence artificielle [Haye77a] [Haye77b] [Wilk72] [Hirs92], qui reposent sur l'analyse syntaxique complète, les bases de sondage, les réseaux sémantiques, l'application de préférences de sélection, etc. Dans d'autres travaux, la syntaxe est combinée avec des informations de collocation fréquentes : [KeSt75], [Dahl88], et [Atki87] combinent des informations de collocation avec des règles pour déterminer, par exemple, la présence ou non de déterminants, pronoms, nom compléments, ainsi que des prépositions, des relations sujet-verbe et verbe-objet, etc.

Les chercheurs ont évité le traitement complexe en utilisant l'analyse superficielle ou partielle. Dans son travail de désambiguïsation sur les noms, [Hear91] segmente le texte en phrases nominales et prépositionnelles et en groupes de verbes, et écarte toutes les autres informations syntaxiques. Il examine les éléments qui se situent à l'intérieur de segments de plus ou moins trois (3) phrases de la cible et combine des preuves syntaxiques avec d'autres types de preuves, telles que la capitalisation. [Yaro93] a déterminé divers comportements basés sur la catégorie syntaxique, par exemple, les verbes dérivent plus d'informations ambiguës de leurs objets que de leurs sujets, les adjectifs dérivent presque toutes les informations ambiguës des noms qu'ils modifient, et les noms sont mieux désambiguïsés par des adjectifs ou des noms directement adjacents. Dans d'autres travaux, les informations syntaxiques font le plus souvent un simple POS, utilisées invariablement en conjonction avec d'autres types d'information [Mcro92] [BrWi94b] [LeMC98].

La catégorie grammaticale du mot ambigu peut également expliquer le besoin de recourir à un contexte symétrique. [Audi03] soutient que, contrairement aux noms et aux adjectifs pour lesquels la majeure partie de l'information levant l'ambiguïté se situe au sein d'un contexte de ± 1 mot autour du mot ambigu, pour les verbes la partie essentielle de l'information se trouve en position +2, voire même +3. Par conséquence, un contexte dissymétrique de la forme -2 +4 serait préférable dans le cas des verbes. [CrEL03] ont élaboré, quant à eux, une méthode qui identifie automatiquement la fenêtre optimale pour chaque phrase contenant une instance du mot ambigu ; ce qui élimine le besoin d'identifier a priori la fenêtre optimale pour un mot donné.

2.5.1.2.4. Le contexte thématique

Le contexte thématique inclut les mots de fond qui co-apparaissent avec un sens donné d'un mot, habituellement dans une fenêtre de plusieurs phrases. Les méthodes reposant sur un contexte d'actualité exploitent la redondance dans un texte, c'est-à-dire l'utilisation répétée de mots sémantiquement liés dans un texte sur un sujet donné. Les travaux impliquant un contexte d'actualité utilisent généralement l'approche du sac de mots, dans laquelle les mots du contexte sont considérés comme un ensemble non ordonné.

L'utilisation du contexte thématique est discutée depuis plusieurs années dans le domaine de la recherche d'information [Anth54] [Salt68]. Des travaux récents du WSD ont exploité le contexte thématique : [Yaro92] utilise une fenêtre de 100 mots, à la fois pour dériver des classes de mots apparentés et comme contexte entourant la cible polysémique, dans ses expériences utilisant le thésaurus de Roget[10]. [Voor95] ont expérimenté plusieurs méthodes statistiques en utilisant une fenêtre de deux phrases ; [LeTV93b] [LeTV93a] ont également exploré le contexte thématique du WSD. [GaCY92c], analysant un contexte de ± 50 mots, indiquent que si les mots les plus proches de la cible contribuent le plus à la désambiguïsation, ils ont amélioré leurs résultats de 86% à 90% en élargissant le contexte de ± 6 (un intervalle typique lorsque seul le micro-contexte est considéré) à ±50 mots autour de la cible. Dans une étude connexe, ils affirment que, pour un discours donné, les mots ambigus sont utilisés dans un sens unique avec une forte probabilité ("un sens par discours") [GaCY92b]. [LeMC98] contestent cette affirmation dans leurs travaux combinant le contexte thématique et le contexte local, ce qui montre que cette combinaison est nécessaires pour obtenir des résultats cohérents entre les mots

[10] Le thésaurus de Roget est un thésaurus de langue anglaise largement utilisé, créé en 1805 par Peter Mark Roget (1779-1869).

polysémiques d'un texte. L'étude de [Yaro93] indique que même si l'information contenue dans une grande fenêtre peut être utilisée pour la désambiguïsation des noms, mais pour les verbes et les adjectifs, la taille de fenêtre doit être beaucoup plus petite. Cela confirme l'affirmation selon laquelle le contexte local et thématique sont nécessaires pour la désambiguïsation, et fait souligne la notion de plus en plus acceptée selon laquelle différentes méthodes de désambiguïsation sont appropriées pour différents types de mots.

Les méthodes utilisant le contexte thématique peuvent être améliorées en divisant le texte analysé en sous-thèmes. La façon la plus évidente de diviser un texte est d'utiliser des sections [BrGY83], mais ce n'est qu'une division grossière ; les sous-thèmes évoluent à l'intérieur des sections, souvent en groupes unifiés de plusieurs paragraphes. Une segmentation automatique des textes en telles unités serait évidemment utile pour les méthodes WSD qui utilisent un contexte thématique. Il a été noté que la répétition de mots à l'intérieur de segments ou de phrases successifs est un indicateur fort de la structure du discours [Skor72] [Morr88] [MoHi91], les méthodes qui exploitent cette observation pour segmenter un texte en sous-thèmes commencent à émerger (voir, par exemple, [Hear94] [Eijk94] [RiSA97]).

2.5.1.2.5. Combinaison des informations du domaine et du contexte local

Quelle que soit l'approche considérée, le contexte local est en général estimé comme une source d'informations plus raffinées que celles liées au contexte thématique et comme un bon indicateur de sens lors de l'utilisation d'un classifieur statistique [LeMC98]. Cependant, bien que les différents types d'informations contextuelles (informations du domaine et du contexte thématique, informations du contexte local) soient habituellement distingués dans la littérature, l'importance de cette distinction n'est pas évidente, comme ne l'est pas non plus l'importance respective de chacun de ces types d'informations sur la sélection des sens. Dans certains travaux, les informations des différents types sont même combinées [LeTV93a] [LeMC98] [Yaro92] [GaCY92c]. Le contexte local est censé davantage contribuer à la désambiguïsation mais les résultats sont néanmoins améliorés en étendant le contexte autour du mot ambigu. Leacock *et al.* (*ibid.*) et Yarowsky (*ibid.*) soutiennent pourtant que l'importance d'une telle combinaison dépend de la catégorie grammaticale du mot ambigu. Le bénéfice pouvant en résulter est plus important dans le cas des noms et moins évident dans le cas des verbes et des modificateurs, pour lesquels la considération du contexte local suffit généralement.

Leacock *et al.* (*ibid.*) examinent le rôle du micro-contexte par rapport au contexte thématique et tentent d'évaluer la contribution de chacun. Leurs résultats indiquent que pour

un classifieur statistique, le micro-contexte est supérieur au contexte thématique en tant qu'indicateur de sens. Cependant, même si une distinction est faite entre le micro-contexte et le contexte thématique dans les travaux actuels sur les WSD, il n'est pas clair que cette distinction soit significative. Il serait peut-être plus utile de considérer les deux comme un continuum et de tenir compte du rôle et de l'importance de l'information contextuelle en fonction de la distance par rapport à la cible.

2.6. Les approches de WSD

Dans la littérature, différentes approches ont été utilisées pour WSD. Les langues riches en ressources telles que l'anglais, le français, l'allemand, et d'autres langues européennes peuvent s'appuyer sur diverses ressources pour la désambiguïsation, tels que les dictionnaires lisibles par machine, thésaurus, ontologies, etc. Cependant, d'autres langes ne disposent pas beaucoup de ressources à l'heure actuelle.

2.6.1. Les approches basées sur les connaissances

2.6.1.1. Les approches basées sur les dictionnaires et les thésaurus

Ces méthodes reposent sur les ressources de connaissances, tels que les dictionnaires lisibles par machine (Machine Readable Dictionaries : MRD), des thésaurus et des ontologies, etc. et elles peuvent utiliser des règles de grammaire pour la désambiguïsation. Au cours des dernières années, de nombreux dictionnaires sont disponibles au format MRD, comme : Oxford English Dictionary (OED[11]), Collins Dictionary (CD[12]), Longman Dictionary of Contemporary English (LDOCE[13]), et les thésaurus qui ajoutent des informations de synonymie comme le thésaurus Roget. Les formats MRD incluent une liste de sens, des définitions (pour tous les sens de mots) et des exemples d'utilisation typiques, tandis qu'un thésaurus ajoute une synonymie explicite entre les sens des mots et le réseau sémantique. L'algorithme de Lesk (Figure 2.1) est l'algorithme fondamental basé sur un dictionnaire pour WSD. C'est l'un des premiers algorithmes développés pour la désambiguïsation sémantique de mots dans un contexte donné. Il a été développé par Michael Lesk en 1986 [Lesk86]. Dans le cadre de cette

[11] Oxford English Dictionary (OED) est un dictionnaire de référence pour la langue anglaise. Il est publié par l'Oxford University Press. La première édition complète, comprenant vingt tomes, est publiée en 1928. Il est depuis régulièrement remis à jour.

[12] Collins Dictionary (CD) est un dictionnaire anglais imprimé et en ligne. Il a été publié par HarperCollins à Glasgow. L'édition du dictionnaire en 1979 avec Patrick Hanks en tant qu'éditeur et Lawrence Urdang en tant que directeur de la rédaction.

[13] Longman Dictionary of Contemporary English (LDOCE) est un dictionnaire anglais imprimé et en ligne. Il a été publié pour la première fois par Longman en 1978. Le dictionnaire est imprimé et en ligne.

méthode, la sélection du sens véhiculé par un mot ambigu se fait en calculant le recouvrement entre les mots inclus dans les définitions des sens du mot et ceux inclus dans les définitions des cooccurrents au sein d'un dictionnaire. Le principal inconvénient de la méthode de Lesk qu'elle repose sur l'appariement exacte entre les mots trouvés dans les définitions. Cette exigence la fait donc dépendre fortement des mots utilisés dans les définitions et la rend très sensible à la présence (ou non) d'un mot au sein de ces définitions. Elle ne lui permet pas, en outre, de capter les relations qui ne sont pas explicitement décrites dans les définitions (Véronis et Ide, 1990). Malgré cet inconvénient, l'idée principale de la méthode de Lesk a été reprise et élaborée dans de nombreux travaux qui ont suivi.

[1] Pour chaque sens s_1 de m_1 et s_2 de m_2

[2] Calculer le Chevauchement (s_1, s_2).

Compter les mots communs entre les gloses des sens s_1 et s_2

[3] Trouver un chevauchement maximisé (s_1, s_2) de s_1 et s_2.

[4] Attribuer le sens s_1 à m_1 et s_2 à m_2.

Figure 2.1: Algorithme de Lesk basé sur un dictionnaire

Étant donné deux mots (m_1, m_2) avec leurs sens définis dans les deux dictionnaires D_1 et D_2 respectivement. Le sens du mot peut être déterminé en utilisant le recouvrement le plus élevé de la définition correspondante en comptant le nombre de mots communs. Ensuite, la paire de sens avec le recouvrement le plus élevé est sélectionnée et un sens attribué au recouvrement maximum pour chaque mot des mots initiaux. Le tableau 1.1 présente un exemple sur l'application des étapes de l'algorithme de Lesk pour désambiguïser les mots dans la paire de mots anglais « Pine » et « Cone » (« pin » et « cône ») [Lesk86].

Pine

1. seven kinds of evergreen tree with needle-shaped leaves

2. waste away through sorrow or illness

3. pine (desperately desire) for something, pine (desperately want) to do something

Cone

1. solid body which narrows to a point

2. Something of this shape, whether solid or hollow

3. fruit of certain evergreen trees (fir, pine)

Figure 2.2: Exemple sur l'application de l'algorithme de Lesk

L'algorithme de « *Lesk adapté* » est l'une des variantes de l'algorithme de Lesk. Il a été introduit par Banerjee et al. en 2002 [BaPe02] . Cet algorithme fondé sur l'approche de désambiguïsation de Lesk, mais au lieu d'utiliser les définitions des dictionnaires traditionnels, sa méthode exploite les informations contenues dans les relations lexicales définies par WordNet. L'algorithme de Lesk repose sur la révélation de recouvrements entre les définitions dictionnairiques de mots voisins au mot à désambiguïser, tandis que celui de Banerjee et Pedersen (*ibid.*) étend les comparaisons aux définitions de mots liés à la fois au mot ambigu et aux mots de son contexte, au sein de WordNet. L'algorithme de « Lesk adapté » prend en compte les hypernymes[14], les hyponymes[15], les holonymes[16], les méronymes[17], les troponymes[18], les relations d'attributs et leurs définitions associées, afin de créer un contexte élargi pour le sens d'un mot donné. La richesse des informations ainsi exploitées améliore la précision de la désambiguïsation [AgEd07].

[14] L'hypernymie et l'hyponymie sont des relations entre mots du lexique caractérisées par le degré de généralité / spécificité. La référence de l'hypernyme (terme plus général) contient celle de tous ses hyponymes (termes plus spécifiques).

[15] Les hyponymes sont inclus dans la classe désignée par ses hypernymes.

[16] L'holonymie est une relation partitive hiérarchisée : M_1 est un holonyme de M_2 si son signifié comprend le signifié de M_2. Par exemple : corps est un holonyme de bras et maison est un holonyme de toit.

[17] La relation inverse est l'holonymie.

[18] La troponymie est la présence d'une relation de « manière » entre deux lexèmes. La notion a été initialement proposée par Christiane Fellbaum et George Miller.

2.6.1.2. Préférences de sélection

Les préférences de sélection font référence au degré de corrélation entre deux catégories linguistiques concomitantes. C'est une ressource utile et polyvalente pour la désambiguïsation du sens des mots. Les préférences de sélection capturent des informations sur les relations possibles entre les catégories de mots et représentent des connaissances de bon sens sur les classes de concepts, par exemple EAT-FOOD (Manger-nourriture), DRINK-LIQUID (boire-liquide). De telles contraintes sémantiques, qui peuvent être utilisées pour éliminer les sens des mots incorrects et non sélectionner seulement les sens qui sont en harmonie avec les règles du sens commun. Par exemple, étant donné la phrase « J'ai mangé un avocat », le sens du plaideur de l'avocat ne correspond pas au contexte ; le verbe mangé nécessite une nourriture comme objet direct.

Bien que les préférences de sélection soient intuitives et nous viennent à l'esprit de manière naturelle, il est difficile de les mettre en pratique pour résoudre le problème de WSD. La raison principale semble être la relation circulaire entre les préférences de sélection et WSD : pour apprendre des contraintes sémantiques précises, il faut connaître le sens des mots impliqué dans une relation de candidat et vice versa, WSD peut être améliorée si des grandes collections de préférences de sélection sont disponibles [AgEd07].

[Resn97a] a exploré comment un modèle statistique de préférences de sélection, qui n'exige ni annotation manuelle des restrictions de sélection ni un apprentissage supervisé, peut être utilisé pour la désambiguïsation. [AgMa01] ont évalué l'application de préférences de sélection mot-à-mot, mot-à-classe et classe-à-classe pour WSD. [StWi01] ont évalué ces applications en tant que caractéristiques de WSD. Les préférences de sélection ont été dérivées dans leur travail en utilisant : (1) le code sémantique LDOCE ; (2) une hiérarchie construite sur mesure pour ces codes qui indique par exemple que les liquides, les gaz et les solides sont toutes sortes d'inanimés ; (3) des relations grammaticales telles que sujet-verbe, verbe-objet et modificateurs de noms identifiés à l'aide d'un analyseur syntaxique peu profond. [McCa03]ont évalué les préférences de sélection acquises automatiquement pour une utilisation dans un système WSD non supervisé. L'utilisation de préférences de sélection pour WSD est une méthode attrayante, en particulier lorsque ces préférences peuvent être apprises sans utiliser de données étiquetés.

[Ye04] a étudié la possibilité d'obtenir de meilleures performances en utilisant les systèmes d'étiquetage de rôles sémantiques de l'état actuel de la technique, Ye (*ibid.*) avait

exploré d'autres moyens d'appliquer les préférences de sélection pour WSD. [ShTM15] ont présenté leur première méthode d'apprentissage de préférences de sélection qui extrait simultanément des connaissances de textes, de vidéos et d'images en utilisant des descriptions d'images et de vidéos pour obtenir des caractéristiques visuelles.

2.6.1.3. Méthodes heuristiques

Ces méthodes utilisent différentes propriétés linguistiques pour trouver le sens du mot ambigu dans un contexte donné. C'est un moyen facile et assez précis de prédire le sens d'un mot sur la base d'heuristiques tirées de propriétés linguistiques observées sur un texte volumineux. Voici trois types d'heuristiques utilisés pour attribuer un sens à certaines catégories de mots.

2.6.1.3.1. Le sens le plus fréquent

Cette méthode consiste à trouver tous les sens probables qu'un mot peut avoir et le sens approprié apparaît plus fréquemment. C'est une méthode très simple et souvent utilisée comme base de référence pour WSD [GaCY92d]. Cependant, cette méthode présente un inconvénient important : la distribution des sens n'est pas toujours disponible et, par conséquent, l'heuristique des sens la plus fréquente ne s'applique qu'aux rares langues pour lesquelles des corpus étiquetés de manière significative sont disponibles. Il existe également une méthode alternative pour rechercher le sens le plus fréquent, qui ne suppose pas la disponibilité de données étiquetées de sens. [MKWC04] montrent comment une mesure de similarité entre diverses sens d'un mot et des mots similaires sur le plan de la distribution peut être utilisée pour déterminer le sens prédominant dans un domaine donné.

2.6.1.3.2. Un sens par discours

Cette approche heuristique a été introduite par Gale et al. [GaCY92b]. Elle affirme qu'un mot tend à conserver sa signification dans toutes ses occurrences dans un discours donné. Il s'agit d'une règle assez forte car elle permet la désambiguïsation automatique de toutes les instances d'un certain mot, étant donné que sa signification est identifiée dans au moins un événement de ce type.

Initialement, le principe « un sens par discours » a été testé sur neuf mots avec une ambiguïté bidirectionnelle dans une expérience réalisée avec cinq sujets. Les sujets ont reçu 82 paires de lignes de concordance et on leur a demandé de déterminer si elles correspondaient ou non au même sens. Dans l'ensemble, ils ont constaté qu'avec une probabilité de 98%, la présence de deux mots dans le même discours aurait le même sens [GaCY92b].

La désambiguïsation se base sur le principe un sens par discours, principe régissant également la méthode proposée par Yarowsky [Yaro95a], selon lequel, les différentes instances du mot dans le texte véhiculent tous le même sens. Dans ce travail, Yarowsky (*ibid.*) a combiné ce principe avec un autre « un sens par collocation ». Ce principe prend en compte le contexte local du mot à désambiguïser, supposé fournir des indices forts et consistants sur le sens du mot, conditionnés par la distance relative, l'ordre et la relation syntaxique. Le mot «collocation» est employé ici dans son sens traditionnel, à savoir, les mots apparaissant au même endroit ou une juxtaposition de mots. Aucune interprétation idiomatique ou non-compositionnelle n'y est impliquée. Les deux méthodes précitées [GaCY92b] et [Yaro95a] visent néanmoins à distinguer les deux sens principaux des mots étudiés, liés à des sujets différents.

Bien que cette hypothèse soit extrêmement vraisemblable pour des mots avec des distinctions de sens approximativement similaires, [Krov98] a expérimenté des mots qui ont plus que deux sens possibles et / ou des distinctions de sens plus fines et a constaté que ces mots ont tendance à avoir plus d'un sens par discours. Ses évaluations ont montré qu'environ 33% des mots des textes se sont avérés avoir plusieurs sens par discours et la précision globale de la désambiguïsation obtenue dans ce cas est donc inférieure à 70%.

2.6.1.3.3. Un sens par collocation

L'heuristique d'un sens par collocation est semblable dans son esprit à l'hypothèse d'un sens par discours mais elle a une portée différente. Il a été introduit par Yarowsky [Yaro93] qui indique qu'un mot tend à garder son sens lorsqu'il est utilisé dans la même colocation. En d'autres termes, les mots voisins fournissent des indices forts et cohérents sur le sens d'un mot cible. Il a également été observé que cet effet est plus fort pour les collocations adjacentes et s'affaiblit à mesure que la distance entre les mots augmente.

Les expériences initiales avec cette hypothèse ont à nouveau considéré des distinctions de sens approximativement similaires, principalement des mots avec une ambiguïté dans les deux sens. Une précision globale de 97% a été observée sur un grand nombre d'exemples annotés à la main. Cette hypothèse est combinée avec le principe un sens par collocation (voir section 2.5.1.2.2).

2.6.1.3.4. Similarité sémantique

Les mots dans un discours doivent avoir un sens pour que le discours soit cohérent [HaHa14]. C'est une propriété naturelle du langage humain et en même temps une des

contraintes les plus puissantes utilisées pour la désambiguïsation automatique des mots. Les mots qui partagent un contexte commun ont généralement une signification proche l'une de l'autre. Par conséquent, les sens appropriés peuvent être sélectionnés en choisissant les significations trouvées dans la plus petite distance sémantique [RMBB89].

Alors que ce type de contrainte sémantique est souvent capable de fournir une unité à un discours entier, sa portée a été généralement limitée à un petit nombre de mots trouvés à proximité immédiate d'un mot cible ou à des mots liés par des dépendances syntaxiques avec le mot cible. Ces méthodes ciblent le contexte local d'un mot donné et ne prennent pas en compte les informations contextuelles supplémentaires trouvées en dehors d'une certaine taille de fenêtre.

Il existe cependant d'autres méthodes qui reposent sur un contexte global et tentent de construire des instances de sens dans tout un texte avec leur portée étendue au-delà d'une petite fenêtre sur les mots cibles. Les chaînes lexicales sont un exemple de telles relations sémantiques établies entre plusieurs mots d'un texte.

À l'instar de l'algorithme de Lesk, ces méthodes de similarité deviennent extrêmement intensives en calcul lorsque plus de deux mots sont impliqués. Cependant, les solutions conçues pour augmenter l'efficacité de l'algorithme de Lesk sont également applicables ici, comme par exemple l'algorithme proposé dans [AgRi96], dans lequel chaque mot ambigu est désambiguïsé individuellement en utilisant une méthode semblable à celle de l'algorithme simplifié de Lesk.

2.6.2. Méthodes supervisées

Ces méthodes sont basées sur des corpus étiquetés (données d'apprentissage) pour WSD. Les informations sont tirées de l'apprentissage sur certains corpus. Un corpus fournit un ensemble d'échantillons qui permet au système de développer des modèles numériques. Les méthodes supervisées reposent sur l'hypothèse que le contexte peut se fournir suffisamment de preuves pour lever toute ambiguïté. Ces méthodes sont sujettes à un nouveau goulot d'acquisition des connaissances car elles reposent sur une quantité importante de corpus étiquetés manuellement pour l'apprentissage, qui sont laborieux et coûteux à créer [LeNC04]. En général, les approches WSD supervisées ont donné de meilleurs résultats que les autres approches.

2.6.2.1. Classifieurs bayésiens naïfs

Un classifieur bayésien naïf [Pede00] [EsMR00] [LeSh04] est une méthode bien connue de la communauté d'apprentissage automatique pour obtenir de bons résultats en matière de désambiguïsation du sens des mots. Cet algorithme est considéré sous approche probabiliste ; celle-ci est une méthode statistique qui estime généralement un ensemble de paramètres probabilistes qui expriment les distributions de probabilité conditionnelles ou conjointes de catégories et de contextes. Il est basé sur le théorème de Baye dans lequel la probabilité conjointe est calculée pour chaque sens S d'un mot M sur les caractéristiques définies (X_1, X_2, ……, X_n) dans un contexte donné.

$$\begin{aligned} \underset{s}{\operatorname{argmax}} P(S|X_1, X_2, \dots, X_n) &= \underset{s}{\operatorname{argmax}} \frac{P(X_1, X_2, \dots, X_n|S)\, P(S)}{P(X_1, X_2, \dots, X_n)} \\ &= \underset{s}{\operatorname{argmax}} P(S) \prod_{I=1}^{n} P(X_i|S) \end{aligned} \qquad \text{(Eq.2.1)}$$

La valeur maximale évaluée à partir de la formule (Eq.2.1) représente le sens le plus approprié dans le contexte, el le nombre de caractéristiques est représenté par n, et la probabilité $P(S)$ est calculée à partir de la fréquence de cooccurrence du sens dans un corpus d'apprentissage et $P(X_i|S)$ est calculée à partir de la caractéristique en présence du sens.

2.6.2.2. Liste de décision

Une liste de décision consiste en un ensemble de règles ordonnées de la forme (valeur-caractéristique, sens, poids). Dans ce contexte, l'algorithme de listes de décision fonctionne comme suit : les données d'apprentissage sont utilisées pour estimer les caractéristiques pondérées par une mesure de « log-likelihood » (logarithme de vraisemblance) [Yaro95a] indiquant la probabilité d'un sens donné à partir d'une valeur de caractéristique particulière. La liste de toutes les règles est triée par valeurs décroissantes de ce poids ; lors du test de nouveaux exemples, la liste de décision est vérifiée et la caractéristique avec le poids le plus élevé qui correspond à l'exemple de test sélectionne le sens du mot.

La formule originale de [Yaro95b] peut être adaptée pour traiter les problèmes de classification comportant plus de deux classes. Dans ce cas, le poids du $sens_k$ lorsque la caractéristique$_i$ apparaît dans le contexte est calculé en tant que logarithme de la probabilité du $sens_k$ sachant la $caractéristique_i$ est divisée par la somme des probabilités des autres sens sachant la $caractéristique_i$ (Eq.2.2).

$$poid(sens_k, caractéristique_i) = \log\left(\frac{p(sens_k|caractéristique_i)}{\sum_{j\neq k} p(sens_j|caractéristique_i)}\right) \qquad \text{(Eq.2.2)}$$

Ces probabilités peuvent être estimées à l'aide de l'estimation de la vraisemblance maximale et d'un lissage pour éviter le problème de division par zéro.

Ces probabilités peuvent être calculées à l'aide de l'estimation de la vraisemblance maximale et d'un lissage afin d'éviter le problème de division par zéro. Il existe de nombreuses approches pour lisser les probabilités. [Chen96] ont proposé une étude complète des différentes techniques de lissage. Pour leurs expériences, [MEMR07] ont adopté une solution très simple : remplacer le dénominateur par 0.1 lorsque la fréquence est nulle.

2.6.2.3. Arbre de décision

L'arbres de décision est une structure de données de l'apprentissage statistique. Son fonctionnement repose sur des heuristiques qui, tout en satisfaisant l'intuition, donnent des résultats remarquables en pratique. Sa structure arborescente le rend également lisible par un être humain.

L'arbre de décision [Pede01] [SGNB14] est utilisé pour désigner les règles de classification dans une arborescence qui divise le jeu de données d'apprentissage de manière récursive. Le nœud interne d'un arbre de décision indique un test qui va être appliqué à une valeur d'entité et chaque branche indique une sortie du test. Lorsqu'un nœud feuille est atteint, le sens du mot est représenté.

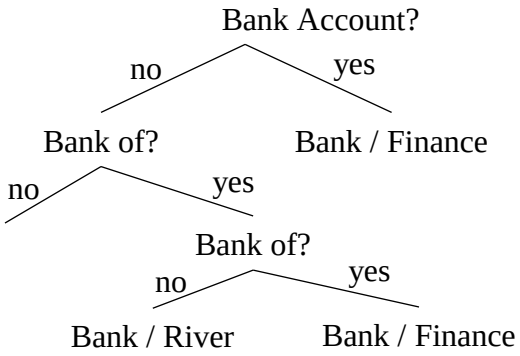

Figure 2.3: Un exemple d'arbre de décision [Navi09]

L'exemple ci-dessous (Figure 2.3) décrit l'utilisation d'arbre de décision pour WSD, dont le sens ambigu du mot « Bank » en anglais est classé dans la phrase « I will be at bank of Narmada River in the afternoon » (Je serai au bord de la rivière Narmada dans l'après-midi).

Dans cet exemple, l'arborescence est créée et parcourue et la sélection est disponible pour cette valeur d'entité.

2.6.2.4. Machines à vecteurs de support (SVM)

Les machines à vecteurs de support ou séparateurs à vaste marge (Support Vector Machine : SVM) a été introduite par Boser et al. [BoGV92]. Depuis lors, cette méthode a acquis une acceptabilité considérable dans la traduction automatique. De nos jours, les SVM ont été appliqués avec succès à un certain nombre de problèmes liés à la reconnaissance de formes en bio-informatique et à la reconnaissance d'images. En ce qui concerne le traitement de texte, les SVM ont obtenu les meilleurs résultats à ce jour en matière de catégorisation de texte [Joac98] et ils sont utilisés dans un nombre croissant de problèmes liés à la NLP, par exemple la fragmentation [KuMa01], l'analyse [Coll04], et WSD [MUUM01] [LeNg02] [EsMR04] [LeNC04].

Les SVM sont basés sur le principe de minimisation des risques structurels de la théorie de l'apprentissage statistique [Vapn98] et, dans leur forme de base, ils apprennent un hyperplan linéaire qui sépare un ensemble d'échantillons positifs d'un ensemble d'échantillons négatifs avec une marge[19] maximale. Ce biais d'apprentissage s'est révélé avoir de bonnes propriétés en termes de bornes de généralisation pour les classifieurs induits.

Le classifieur linéaire est défini par deux éléments : un vecteur de pondération w (avec une composante pour chaque entité) et un biais b qui représente la distance de l'hyperplan à l'origine (Figure 2.4). La règle de classification affecte +1 ou –1 à un nouvel exemple x comme suit :

$$h(x) = \begin{cases} +1, & < w, x > + b \geq 0 \\ -1, & autrement \end{cases} \qquad \text{(Eq.2.3)}$$

Les SVMs sont des classifieurs qui reposent sur deux idées, qui permettent de traiter des problèmes de discrimination non linéaire, et de reformuler le problème de classement comme un problème d'optimisation quadratique convexe[20].

[19] La marge γ est définie par la distance de l'hyperplan au plus proche des exemples positifs et négatifs, la marge γ est définie par la distance entre la frontière de séparation et les échantillons les plus proches

[20] Un problème d'optimisation quadratique est un problème d'optimisation dans lequel on minimise (ou maximise) une fonction quadratique sur un polyèdre convexe (Un objet géométrique est dit convexe lorsque, chaque fois qu'on y prend deux points A et B, le segment [A, B] qui les joint y est entièrement contenu). Les contraintes peuvent donc être décrites par des fonctions linéaires (on devrait dire affines). L'optimisation quadratique (OQ) est la discipline qui étudie ces problèmes. L'optimisation linéaire peut être vue comme un cas particulier de l'optimisation quadratique.

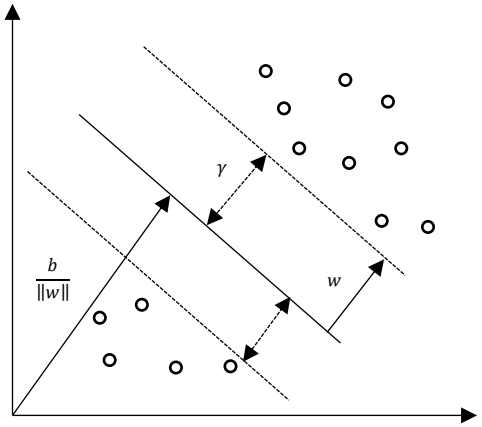

Figure 2.4: Exemple de marge γ et d'hyperplan <w,x>+ b

L'apprentissage de l'hyperplan de la marge maximale (w, b) peut être simplement défini comme un problème d'optimisation quadratique convexe avec une solution unique, consistant en forme primaire : minimiser $\|w\|$ en fonction des contraintes (un pour chaque exemple d'apprentissage) : $y_i\ (< w, x_i > + b) \geq 1,\ \forall\ 1 \leq i \leq N$ où N est le nombre d'exemples d'apprentissage.

Parfois, les exemples d'apprentissage ne sont pas séparables linéairement ou, simplement, il n'est pas souhaitable d'obtenir un hyperplan parfait. Dans ces cas, il est préférable de permettre certaines erreurs dans l'ensemble d'apprentissage afin de maintenir un « meilleur » *hyperplan solution* (Figure 2.5). Ceci est réalisé par une variante du problème d'optimisation, appelée *marge molle*, dans laquelle la contribution à la fonction objective de la maximisation de la marge et des erreurs d'apprentissage peut être compensée par l'utilisation d'un paramètre appelé C. Ce paramètre affecte les variables d'écart (Slack Variables) ϑ_i dans la fonction. Ce problème peut être formulé comme suit :

$$minimiser : \frac{1}{2} < w, w > + C \sum\nolimits_i^N \vartheta_i \qquad \text{(Eq.2.4)}$$

$$En\ fonction\ de : y_i(< w, x_i > + b) \geq 1 - \vartheta_i, \vartheta_i \geq 0, \forall\ 1 \leq i \leq N$$

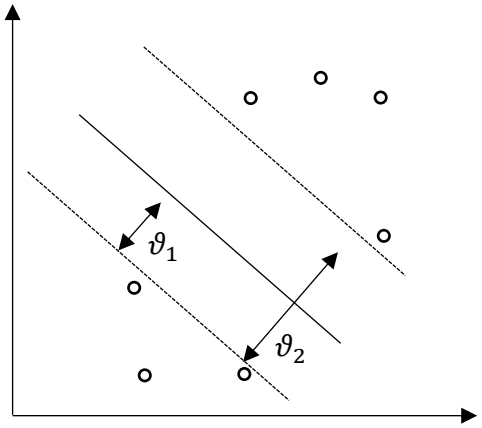

Figure 2.5: Exemple de marge molle avec des variables d'écart ϑ_i [BaVC06]

2.6.2.5. Réseaux de neurones

Il s'agit d'une interconnexion de neurones artificiels utilisant un modèle informatique pour le traitement de données basé sur des approches connexionnistes. Ces approches ont été suggérées par Cottrell et al., [CoSm83], Waltz et al., [WaPo85] pour la désambiguïsation du sens des mots. Le modèle de réseau neuronal est constitué de réseaux dans lesquels les nœuds représentent des mots reliés par des lignes dirigées : le mot active les concepts auxquels il est lié sémantiquement et vice versa. De plus, les liens inhibiteurs « latéraux » relient généralement les sens concurrents d'un mot donné. Initialement, les nœuds correspondant aux mots de la phrase à analyser sont activés. Ces mots activent leurs voisins dans le cycle suivant, tour à tour, ces voisins activent leurs voisins immédiats, etc. Après un certain nombre de cycles, le réseau se stabilise dans un état dans lequel un sens pour chaque mot entré est plus activé que les autres, en utilisant un processus de relaxation parallèle et analogique. Les approches de réseau de neurones, pour WSD, semblent en mesure de capturer la plupart de ce qui ne peut pas être traité par des stratégies de chevauchement telles que celles de Lesk. Cependant, les réseaux utilisés jusqu'à présent dans les expériences, à notre connaissance, sont codés à la main et sont donc nécessairement très petits (tout au plus quelques dizaines de mots et de concepts). En raison du manque de données réalistes, il n'est pas clair que les mêmes modèles de réseaux de neurones seront mis à l'échelle pour des applications réalistes.

2.6.2.6. Apprentissage basé sur des exemples

L'algorithme de K plus proche voisin (k-Nearest Neighbor : kNN) est un exemple d'apprentissage basé sur des exemples. Cet algorithme est la meilleure option pour WSD [Ng97]. [DaBZ99] soutiennent que les méthodes basées sur des exemples ont tendance à être meilleures dans les problèmes de NLP car elles traitent les exceptions. En kNN, tous les

exemples sont stockés en mémoire pendant l'entraînement et la classification d'un nouvel exemple est basée sur les sens des k exemples stockés les plus similaires. Pour obtenir l'ensemble des voisins les plus proches, l'exemple à classer $x = (x_1, x_2, \dots, x_m)$ est comparé à chaque exemple stocké $x^i = \left(x_1^i, x_2^i, \dots, x_m^i\right)$, et la distance entre eux est calculé. La métrique la plus basique (également appelée distance de Hamming) est définie comme suit :

$$\Delta\left(x, x^i\right) = \sum_{j=1}^{m} w_j \delta(x_j, x_j^i) \qquad \text{(Eq.2.5)}$$

Où w_j est le poids de la j[ième] caractéristique et $\delta(x_j, x_j^i)$ est la distance entre deux valeurs, égale à 0 si $x_j = x_j^i$ et à 1 sinon.

Les chercheurs ont utilisé la distance de Hamming pour mesurer la proximité et le ratio de gain [Quin14] pour estimer le poids des caractéristiques. Pour k valeurs supérieures à 1, le sens résultant est le sens pondéré de la majorité des k plus proches voisins, où chaque exemple exprime son sens avec une force proportionnelle à sa proximité avec l'exemple de test. L'algorithme est répété un certain nombre de fois en utilisant un nombre différent de k (voisins les plus proches, c'est-à-dire k = 1, 3, 5, 7, 10, 15, 20 et 25). Les résultats correspondant au meilleur choix pour chaque mot sont utilisés.

2.6.2.7. Méthode statistique (modèle n-gramme)

Un n-gramme est simplement une séquence de n mots successifs avec leur nombre, c'est-à-dire le nombre d'occurrences dans les données d'apprentissage. Pour des raisons de calcul, nous appliquons des hypothèses de Markov selon lesquelles le mot actuel ne dépend pas de l'historique complet du mot, mais tout au plus des derniers mots. Le nombre de mots dans le contexte local de mots ambigus constitue une fenêtre et la taille de la fenêtre (c'est-à-dire le nombre de mots à prendre en compte) est importante. Au lieu de générer un modèle d'ordre supérieur de n-gramme, qui prend beaucoup de temps, et est difficile à créer et à gérer et nécessite bien sûr beaucoup de données pour obtenir des résultats significatifs, nous pouvons utiliser la combinaison du modèle d'ordre inférieur n-gramme [JuMa14].

2.6.3. Méthodes non supervisées

Ces méthodes sont basées sur des corpus non étiquetés [BrWi94a] [BrWi94b] [PeBr97] [PeBr98] [Schü98] [Véro03]. Ceux-ci doivent être formés, avant d'être utilises, sur des mots ambigus. L'avantage d'utiliser ce type de corpus est d'éviter l'étiquetage, qui est un processus

longue et coûteux (c'est ce qu'on appelle le goulot d'étranglement de l'acquisition de connaissances). Les connaissances nécessaires à la désambiguïsation sont automatiquement identifiées dans les textes traités. Les sens possibles des mots ambigus sont repérés dans les textes en regroupant les instances des mots sur la base de traits contextuels divers.

L'analyse du contexte, effectuée pour la détermination des sens des mots ambigus, permet le repérage des traits avec lesquels le contexte de nouvelles instances des mots sera comparé, par la suite, pour la désambiguïsation. Dans la méthode de Schütze (*ibid.*), par exemple, où les sens lexicaux correspondent à des clusters de vecteurs contextuels, la désambiguïsation d'une nouvelle instance d'un mot s'opère en comparant le vecteur construit pour le nouveau contexte avec la centroïde de chaque cluster (la moyenne de ses éléments) et en sélectionnant ensuite le cluster dont la centroïde est la plus proche du vecteur contextuel. Le cluster retenu correspond au sens du mot dans le nouveau contexte.

En revanche, dans le travail de Pedersen et Bruce (*ibid.*), l'étape d'acquisition de sens coïncide avec celle de la désambiguïsation. Le processus de désambiguïsation est appliqué sur un corpus sémantiquement étiqueté à des fins d'évaluation. Et les étiquettes sémantiques ne sont pas utilisées lors de l'apprentissage (qui est non supervisé) mais servent à l'évaluation des groupes de sens générés, mis en correspondance avec les étiquettes.

Les différentes méthodes utilisées dans une approche non supervisée sont le clustering de contexte, et le clustering de mots et graphes de cooccurrence [SiMi07] [JuMa14]. Les limitations de cette approche sont les suivantes :

- Elle ne convient pas aux situations à grande échelle ;
- Les instances dans les données d'apprentissage peuvent ne pas attribuer le sens correct ;
- L'apprentissage de clusters hétérogènes et le nombre de clusters peuvent différer du nombre de sens du mot cible.

Toutefois, d'autres chercheurs ont constaté que les performances de cette méthode ont été inférieures à celles d'autres méthodes [CDBB09].

2.6.3.1. Le clustering de contexte

Dans la méthode de clustering de contextes, chaque occurrence du mot cible dans le corpus est représentée en tant que vecteur de contexte. Ces vecteurs sont regroupés en clusters pour identifier le sens du mot cible. L'espace vectoriel est constitué de mots en tant que

dimensions et tous les sens possibles du mot sont maintenus dans le vecteur. Le cosinus entre deux vecteurs représente la similarité entre ces vecteurs. La matrice de cooccurrence est développée en regroupant les vecteurs. Parfois, la matrice rencontre le problème de la haute dimensionnalité qui peut être résolu en fusionnant les dimensions ayant le même sens du mot [AgEd07]. L'avantage de cette approche est qu'une grande quantité de données d'apprentissage annotées manuellement n'est pas nécessaire, tandis que son inconvénient est que les données d'apprentissage sont requises pour chaque mot devant lever l'ambiguïté. Divers algorithmes de clustering sont disponibles dont quelques-uns sont expliqués comme suit :

1. Discrimination de groupe de contexte [Schü98] : Chaque occurrence du mot ambigu est regroupée dans un cluster de sens. Ce regroupement est basé sur la similitude de contexte dans lequel le mot est apparu. Le cosinus entre les vecteurs donne la similarité entre les vecteurs et le clustering est effectué à l'aide de L'algorithme d'espérance-maximisation[21] [DeLR77] [JaMF99] [WFHP16].
2. Méthode de clustering par agglomération [PeBr97] : Les algorithmes d'agglomération placent au préalable chaque instance dans un cluster séparé et fusionnent ensuite une paire de clusters à chaque itération, formant ainsi des clusters de plus en plus grands, jusqu'à ce qu'il atteigne une valeur de seuil d'arrêt.

2.6.3.2. Le clustering de mots

La méthode de clustering de contextes [AgEd07] [Lin98a] [KaAu13] est basée sur des techniques de regroupement dans lesquelles les premiers vecteurs de contexte sont créés avant d'être regroupés en cluster afin d'identifier le sens du mot. Cette méthode utilise l'espace vectoriel comme espace de mots et ses dimensions sont des mots. Dans cette méthode, les occurrences d'un mot dans un corpus seront désignées comme vecteur et la fréquence de ce mot sera compté dans son contexte. Après, une matrice de cooccurrence est créée et des mesures de similarité sont appliquées. Ensuite, la discrimination est effectuée en utilisant n'importe quelle technique de clustering.

Des algorithmes de clustering de mots sont appliqués pour discriminer les sens. Soit M la liste des mots similaires classés par degré de similitude avec le mot ambigu m_0. Un arbre de similarité T est initialement créé et qui consiste en un nœud unique m_0. Ensuite, pour chaque $i \in \{1,2,\dots,k\}, m_i \in M$ est ajouté en tant que fils de m_j dans l'arbre T, de sorte que m_j est le

[21] L'algorithme d'espérance-maximisation est caractérisé par de meilleures propriétés de convergence que les autres algorithmes, tout en permettant, lui aussi, l'appartenance partielle à des clusters.

mot le plus similaire à m_i entre $(m_0, m_1, \dots, m_{i-1})$. Après une étape d'élagage, chaque sous arbre enraciné à m_0 est considéré comme un sens distinct de m_0. Dans une approche ultérieure, appelée algorithme de clustering par comité (the clustering by committee, CBC), une méthode de classification par mot différente sera proposée. Pour chaque mot cible, un ensemble de mots similaires a été calculé comme ci-dessus. Pour calculer la similarité, encore une fois, chaque mot est représenté par un vecteur de caractéristiques, chaque caractéristique étant l'expression d'un contexte syntaxique dans lequel le mot apparaît.

Étant donné un ensemble de mots cibles (par exemple, tous ceux apparaissant dans un corpus), une matrice de similarité S est construite de telle sorte que S_{ij} contienne la similarité par paires entre les mots m_i et m_j.

Comme deuxième étape, une procédure récursive est appliquée pour déterminer des ensembles de clusters (appelés comités) d'un ensemble de mots E. À cette fin, une technique de clustring standard, à savoir le clustering de liaisons moyennes, est utilisée. À chaque étape, les mots résiduels qui ne sont pas couverts par aucun comité (c.-à-d. qui ne sont pas assez similaires à la centroïde -la moyenne de ses éléments- de chaque comité) sont identifiés. Des tentatives récursives sont faites pour découvrir plus de comités à partir de mots résiduels. A noter que, comme ci-dessus, les comités confondent les sens car chaque mot appartient à un seul comité.

Enfin, en tant qu'étape de discrimination de sens, chaque mot cible $m \in E$, également représenté comme un vecteur de caractéristiques, est attribué de manière itérative à son groupe le plus similaire, en fonction de sa similarité avec la centroïde de chaque comité. À titre de mesure de discrimination fondée sur le sens, chaque mot cible m, représenté de nouveau comme un vecteur de caractéristiques, est attribué de façon itérative à son groupe le plus semblable, en fonction de sa similitude avec le centroïde de chaque comité. Une fois qu'un mot m est attribué à un comité c, les caractéristiques qui se croisent entre m et les éléments de c sont supprimées de la représentation de m, afin de permettre l'identification de sens moins fréquents du même mot lors d'une itération ultérieure.

2.6.3.3. Graphes de cooccurrence

Des approches basées sur des graphes ont été utilisées récemment pour la désambiguïsation avec un certain succès. Dans un graphe de cooccurrence, l'ensemble des sommets V est constitué de mots apparaissant dans le texte et l'ensemble des arêtes E donne la connexion entre les mots cooccurrents dans le même contexte.

L'approche HyperLex [Véro04] a représenté les mots du paragraphe par les nœuds d'un graphe et chaque deux mots figurants dans le même paragraphe par le bord du graphe dont chaque bord prend un poids correspondant à la fréquence à laquelle les mots reliés par le bord se cooccurrent ensemble.

Le poids peut être représenté par :

$$M_{mn} = 1 - Max\{P(M_m/M_n), P(M_n/M_m)\} \quad \text{(Eq.2.6)}$$

Où : $P(M_m/M_n)$ représente $Fréquence_{mn}/Fréquence_n$, et $Fréquence_{mn}$ est la fréquence à laquelle les mots $M_m$et $M_n$cooccurrent et $Fréquence_n$ représente la fréquence à laquelle M_n apparaît dans le contexte. Les mots ayant une fréquence de cooccurrence plus élevée auront un poids proche de zéro (0) et les mots qui cooccurrent sous une forme rare auront un poids proche de un (1).

Un autre algorithme basé sur les graphes pour dériver les sens des mots est l'algorithme PageRank [AgEd07], largement utilisé dans les moteurs de recherche Google. Cet algorithme PageRank peut être utilisé pour estimer l'importance d'objets dont les relations peuvent être décrites par un graphique. [MiTF04] ont exploré l'applicabilité du PageRank aux réseaux sémantiques et montré que de tels algorithmes de classement basés sur des graphes peuvent être utilisés avec succès dans des applications de NLP. Les chercheurs (*ibid.*) ont proposé et expérimenté un nouvel algorithme de désambiguïsation des mots basé sur la connaissance et non supervisé, qui parvient à identifier le sens de tous les mots dans un texte ouvert avec une précision nettement supérieure à celle des autres algorithmes, basés sur la connaissance, proposés précédemment.

2.6.4. Méthodes semi-supervisées

Les méthodes semi-supervisées sont devenues un domaine de recherche actif dans le domaine de la désambiguïsation des sens des mots. Cela nécessite une estimation de la fonction sur des données non étiquetées avec peu de données étiquetées. Cette approche est inspirée par le fait que les données étiquetées sont souvent coûteuses à générer, alors que les données non étiquetées ne le sont généralement pas. Cependant, le grand défi est de savoir comment utiliser des données mixtes (étiquetées / non étiquetées). Cette méthode nécessite également moins d'effort humain, car des données non étiquetées sont disponibles en abondance et sous forme massive pour lever les ambiguïtés des mots polysémiques.

Une approche répandue est l'amorçage (bootstrapping). Ces méthodes impliquent une phase d'apprentissage (ou d'entraînement) sur un petit ensemble d'instances de mots désambiguïsées et étiquetées manuellement du point de vue sémantique [ZhGo09].

2.6.4.1. Méthodes d'amorçage (bootstrapping)

Cette méthodes impliquent une phase d'apprentissage (ou d'entraînement) sur un petit ensemble d'instances de mots désambiguïsées et étiquetées manuellement du point de vue sémantique [ZhGo09].

Steven Abney [Abne02] définit l'amorçage comme un problème dans lequel la tâche consiste à induire un classifieur à partir d'un petit ensemble de données étiquetées et d'un grand ensemble de données non étiquetées. En principe, l'amorçage est une approche très intéressante pour la désambiguïsation en raison de la disponibilité d'un grand nombre de données non étiquetées. L'un des premiers algorithmes d'amorçage appliqués dans TAL est celui de Yarowsky [Yaro95a] qui a été appliqué à la WSD. Les listes de décisions dans cet algorithme ont été utilisées comme base d'apprentissage supervisée. L'entrée de l'algorithme de Yarowsky (*ibid.*) consiste en un ensemble d'exemples étiquetés, également appelés « graines – Seeds » et un ensemble d'exemples non étiquetés, représentant généralement environ 90% du total. L'algorithme est un processus itératif dans lequel un apprenant de la liste de décisions est construit avec le corpus de graines et appliqué aux données non étiquetées. Lors de la prochaine itération, l'algorithme obtient les règles des graines plus que celles de la meilleure confiance acquise à partir de l'ensemble non étiqueté et un nouvel apprenant est construit. Le processus est répété jusqu'à atteindre certains paramètres d'entraînement. L'un des inconvénients de l'amorçage est la justification théorique du processus d'apprentissage [BaVC06].

2.7. Évaluation

Parmi les études citées précédemment, il est évident qu'il est très difficile de comparer un ensemble de résultats, et par conséquent une méthode à une autre. Le manque de comparabilité résulte de différences substantielles dans les conditions de test d'une étude à l'autre. Par exemple, différents types de textes sont impliqués, y compris des textes purement techniques ou spécifiques à un domaine où l'utilisation du sens est limitée, par rapport aux textes généraux dans lesquels l'utilisation du sens peut être plus variable. Il a été noté que dans un corpus communément utilisé, certains sens de mots de test typiques sont totalement absents. Lorsque différents corpus contenant différents inventaires de sens et des niveaux de fréquence

très différents pour un mot et / ou un sens donné sont utilisés, il devient inutile de tenter de comparer les résultats.

Les mots de test diffèrent d'une étude à l'autre, non seulement des mots dont l'attribution à des sens clairement distincts varie considérablement ou qui présentent des degrés d'ambiguïté très différents (par exemple, « Avocat » le fruit contre « Avocat » le plaideur), mais aussi des mots dans différentes parties du discours et des mots qui apparaissent plus fréquemment dans des usages métaphoriques, métonymiques, etc. En outre, les critères d'évaluation de l'exactitude de l'attribution de sens varient. Différentes études utilisent différents degrés de granularité des sens, allant de l'identification des homographes aux distinctions fines des sens. De plus, les moyens par lesquels une attribution de sens correcte est finalement jugée sont généralement peu clairs. Les juges humains doivent en décider définitivement, mais le manque d'accord entre eux est bien attestée. [AmWh79] indiquent que, même s'il existe une cohérence raisonnable dans l'attribution de sens lors d'affectations de sens successives (84%), l'accord est nettement plus faible entre les experts. [Ahls95] signal un accord entre 63,3 et 90,2% des participants sur son questionnaire d'ambiguïté ; pendant une opération d'attribution de sens en ligne dans un grand corpus. L'accord entre participants est bien moindre, voire pire que le hasard ([Ahls92] [Ahls93] [AhLo93]. [Jorg90] a constaté que le degré de concordance de son expérience utilisant les données du corpus Brown[22] était d'environ 68%.

La difficulté de comparer les résultats de la recherche sur le WSD est récemment devenue une préoccupation au sein de la communauté de ce domaine de recherche et des efforts sont en cours pour développer des stratégies d'évaluation du WSD. Gale et al. (1992b) ont tenté d'établir des limites inférieures et supérieures pour évaluer la performance des systèmes WSD ; leur proposition de résoudre le problème de l'accord entre juges humains en vue d'établir une limite supérieure constitue un point de départ, mais celui-ci n'a pas été largement discuté ni mis en œuvre. Une discussion lors d'un atelier sponsorisé par le groupe d'intérêt spécial ACL sur le lexique (SIGLEX) sur « l'évaluation des tagueurs sémantiques automatiques » a suscité la formation d'un effort d'évaluation de la WSD (SENSEVAL[23]).

[22] Le corpus standard de l'Université Brown a été élaboré dans les années 1960 par Henry Kučera et W. Nelson Francis, en tant que corpus général (collection de textes) dans le domaine de la linguistique. Il contient 500 échantillons de texte en langue anglaise, totalisant environ un million de mots, compilés à partir d'œuvres publiées aux États-Unis en 1961.

[23] La première édition de SENSEVAL a eu lieu à l'été 1998 en anglais, français et italien et a abouti à la tenue d'un atelier au Herstmonceux Castle, Sussex, en Angleterre, du 2 au 4 septembre. Senseval est l'organisation internationale spécialisée à l'évaluation des systèmes WSD. Sa mission est d'organiser et de mener des activités d'évaluation des systèmes WSD en ce qui concerne différents mots, différents aspects de la langue et différentes

Comme on l'a vu plus haut, le WSD n'est pas une fin en soi, mais plutôt une « tâche intermédiaire » qui contribue à une tâche globale telle que la recherche d'informations, la traduction automatique, etc. Cela ouvre la possibilité de deux types d'évaluation pour le travail de WSD (en utilisant une terminologie empruntée à la biologie) : l'évaluation in vitro, où les systèmes WSD sont testés indépendamment d'une application donnée, en utilisant des points de repère spécialement conçus ; et l'évaluation in vivo, où, plutôt que d'être évalués isolément, les résultats sont évalués en fonction de leur contribution à la performance globale d'un système conçu pour une application particulière (par exemple, la recherche d'information).

2.7.1. Évaluation in vitro

L'évaluation in vitro (également appelé évaluation déclarative ou évaluation de performance [ArSH93] [BiCP94] [HiTh97]), malgré son caractère artificiel, permet d'examiner de près les problèmes rencontrés dans une tâche donnée. Dans sa forme la plus élémentaire, ce type d'évaluation implique la comparaison de la sortie d'un système pour une entrée donnée, en utilisant des mesures telles que la précision et le rappel. SENSEVAL envisage actuellement ce type d'évaluation pour les résultats WSD. De manière alternative, l'évaluation in vitro peut porter sur l'étude du comportement et des performances des systèmes sur une série de tests représentant l'éventail des problèmes linguistiques susceptibles de survenir lors de WSD (évaluation typologique ; évaluation diagnostique [ArSH93] [HiTh97]). Une compréhension beaucoup plus approfondie des facteurs impliqués dans la tâche de désambiguïsation est nécessaire avant de pouvoir concevoir des séries de tests appropriées pour l'évaluation typologique des résultats de WSD. Des questions de base telles que le rôle d'une partie du discours dans le WSD, le traitement de la métaphore, la métonymie, etc. dans l'évaluation, comment traiter des mots de degrés et de types de polysémie différents, etc., doivent d'abord avoir une réponse. SENSEVAL nous rapproche probablement de cette compréhension, dans la mesure où, il force la prise en compte de ce qui peut être considéré de manière significative comme une distinction de sens et fournit une mesure de la distance entre la performance des systèmes actuels et une norme prédéfinie.

L'évaluation in vitro envisagée par SENSEVAL exige la création d'un corpus de référence étiqueté manuellement, contenant un ensemble convenu de distinctions entre les sens. Les difficultés pour parvenir à un accord de sens, même entre les experts, ont déjà été décrites.

langues. Son objectif sous-jacent est d'approfondir la compréhension de la sémantique lexicale et de la polysémie. (Voir http://web.eecs.umich.edu/~mihalcea/senseval/overview.html).

Actuellement, la meilleure source apparente de distinctions de sens est WordNet, bien que les problèmes d'utilisation de telles ressources soient bien connus et que leur utilisation ne résolve pas les problèmes de marquage sémantique plus complexe qui va au-delà des distinctions typiques faites dans les dictionnaires et les thésaurus.

[Resn97b] soulignent également qu'une évaluation binaire (correcte / incorrecte) pour WSD n'est pas suffisante et proposent que les erreurs soient pénalisées selon une matrice de distance entre les sens basée sur une organisation hiérarchique. Par exemple, le fait de ne pas identifier les homographes de « bank » (banque en français qui apparaîtrait plus haut dans la hiérarchie) serait pénalisé plus sévèrement que de ne pas distinguer « bank » en tant qu'institution, par opposition à « bank » en tant que bâtiment (ce qui apparaîtrait plus bas dans la hiérarchie). Malgré l'avantage évident de cette approche, elle se heurte au même problème, à savoir l'absence d'une hiérarchie des sens établie et convenue. Resnik et al (*ibid.*), qui étaient conscients de ce problème, suggèrent de créer la matrice de distance des sens sur la base de résultats obtenus en psychologie expérimentale [MiCh91] [Resn95], et même en ignorant le coût de la création d'une telle matrice, la littérature psycholinguistique a clairement montré que ces résultats sont fortement influencés par les conditions expérimentales et la tâche imposée aux sujets [Tabo89] [Tabo91] [RaMo91]. En outre, il n'est pas clair que les données psycholinguistiques puissent être utiles dans le cadre du WSD axé sur une utilisation pratique dans les systèmes de NLP.

En général, l'évaluation de WSD se heurte à des difficultés liées à des critères similaires à ceux faisant face à d'autres tâches telles que le marquage partiel du discours, en raison de la nature insaisissable des distinctions sémantiques.

2.7.2. Évaluation in vivo

Une autre approche de l'évaluation consiste à examiner les résultats dans la mesure où ils contribuent à la performance globale d'une application particulière telle que la traduction automatique, la recherche d'informations, la reconnaissance de la parole, etc. Cette approche (également appelée évaluation de l'adéquation [HiTh97] ; évaluation opérationnelle [ArSH93]), bien qu'elle ne garantisse pas l'applicabilité générale d'une méthode ni ne contribue à une compréhension détaillée des problèmes, ne requiert pas un accord sur les distinctions de sens ou la création d'un corpus pré-étiqueté. Seul le résultat final est pris en compte, soumis à une évaluation adaptée à la tâche à accomplir.

Les méthodes de WSD ont été évoluées en grande partie indépendamment d'applications particulières. Des efforts évidents pour intégrer les méthodes WSD dans des applications du domaine de la recherche d'informations, mais les résultats sont ambigus : [KrCr92] ne signalent qu'une légère amélioration de la recherche utilisant les méthodes WSD ; [Voor93] [Sand94] indiquent que la recherche se dégrade si la désambiguïsation n'est pas suffisamment précise. Par ailleurs, [ScPe95] montrent une nette amélioration de l'extraction (14,4%) en utilisant une méthode combinant recherche par mot et recherche par sens.

Il reste à savoir dans quelle mesure WSD peut améliorer les résultats dans des applications particulières. Toutefois, si le sens est en grande partie fonction de l'utilisation, il se peut que la seule évaluation pertinente des résultats de la WSD soit réalisable dans le contexte de tâches spécifiques.

2.8. Conclusion

La recherche d'information sémantique (RIS) est un domaine de recherche en pleine croissance et fait partie de l'Intelligence Artificielle (IA). Et avec l'avènement d'internet, RIS est devenu une nécessité pour retrouver l'information pertinente, à des requêtes données, d'une quantité massive et croissante des informations stockées.

Toutes les langues naturelles (comme l'anglais, l'arabe, le français, etc.) utilisent des mots qui ont plusieurs significations. Ainsi, le problème de la désambiguïsation du sens des mots (WSD) est apparu, à savoir comment sélectionner le sens correct ou voulu d'un mot donné dans un contexte donné. La résolution de ces problèmes est très importante pour une recherche automatique d'informations pertinentes.

Les méthodes de WSD sont un domaine de recherche intense de nos jours et WSD trouve des applications dans de nombreux domaines variés tels que la traduction automatique, la récupération d'informations, l'extraction d'informations, le répondeur automatique, la reconnaissance vocale, la résumé automatique et la synthèse vocale, etc.

Les principales approches utilisées dans WSD sont (a) les méthodes basées sur la connaissance, (b) les méthodes supervisées, (c) les méthodes non supervisées et (d) les méthodes semi-supervisées. Les méthodes basées sur la connaissance utilisent des dictionnaires lisibles par machine (MRD), thésaurus et ontologies, etc.

Les méthodes supervisées utilisent des données d'apprentissage structurées à partir d'un corpus d'entraînement étiqueté. Cette méthode est basée sur le principe que le sens prévu soit

déterminé en fonction du contexte, cela donne des résultats très précis. Cependant, la construction de tel corpus et son utilisation prennent du temps.

Les méthodes non supervisées utilisent des données non étiquetées pour l'apprentissage. Il est plus facile et moins coûteux à construire et à utiliser. Les mots proches sont utilisés pour créer des groupes de sens.

Les méthodes semi-supervisées utilisent une petite quantité de données annotées et une grande partie des données non annotées.

Certaines recherches sur le WSD dans la langue arabe ont été effectuées par différents chercheurs. Cependant, ces travaux demeurent principalement limités à cause du manque de ressources linguistiques. Le chapitre suivant de la thèse présente brièvement un état de l'art sur les différentes approches utilisées dans la recherche d'information sémantique et plus précisément le WSD et les caractéristiques de la ressource lexicale « le dictionnaire de la langue arabe contemporaine » qui nous allons l'utiliser dans nos travaux.

Chapitre 3 :

Les méthodes à base de connaissances appliquées à l'arabe et la ressource lexicale « le dictionnaire de la langue arabe contemporaine »

3. LES METHODES A BASE DE CONNAISSANCES APPLIQUEES A L'ARABE ET LA RESSOURCE LEXICALE « LE DICTIONNAIRE DE LA LANGUE ARABE CONTEMPORAINE »

3.1. Introduction

La désambiguïsation sémantique de mots (WSD) est une tâche indispensable à la bonne performance de plusieurs applications, par exemple les applications de recherche d'information, d'extraction d'information, de traduction automatique, de fouille de textes, de simplification lexicale de textes, etc. [IdVe98], qui nécessitent un niveau de compréhension du texte, mais à des degrés divers, comme les applications de traduction ont besoin d'un excellent degré de désambiguïsation par rapport aux applications de RI.

Il existe différentes approches de désambiguïsation, comme nous l'avons vue dans le chapitre précèdent (chapitre 2), d'une part, les approches supervisées nécessitant des corpus d'entraînement étiquetés manuellement qui demandent beaucoup de temps et sont très coûteux, d'autre part, des approches non-supervisées. Ces dernières sont intéressantes, car elles n'utilisent pas de corpus étiquetés. Les approches non-supervisées elles-mêmes se divisent en deux : des approches purement non supervisées classiques qui exploitent seulement les motifs présents dans les données ; et des approches à base de connaissances qui utilisent des ressources lexicales et exploitent des connaissances linguistiques.

Dans les approches à base de connaissances, on utilise des ressources sémantiques externes comme les dictionnaires, les thésaurus ou les ontologies et on exploite des connaissances linguistiques pour construire une structure représentant respectivement le contenu non ambigu de documents et de requêtes. Cette structure représentative est le résultat d'une procédure d'indexation sémantique. Plusieurs travaux ont été réalisés à cet égard. Alors que la plupart de ces travaux concernaient les langues latines [IdVé98] [Navi09], peu de travaux sur l'arabe [LBEE13] [ADAC16a]. La langue arabe, caractérisée par une morphologie très complexe, souffre du manque de ressources linguistiques puissantes et structurées, d'outils informatiques efficaces d'annotation et d'évaluation, d'études rigoureuses et du choix de ressources linguistiques et de comment exploiter ces ressources pour désambiguïser.

Dans ce chapitre nous essayerons d'aborder les études les plus intéressantes et les plus récentes sur l'utilisation de l'approche à base de connaissance en langue arabe, et de définir et de clarifier les caractéristiques du dictionnaire « la Langue arabe Contemporaine », lequel nous allons utiliser dans notre étude.

3.2. L'approche à base de connaissance dans les études des WSD en langue arabe

Nous commençons d'abord par une brève présentation des principaux travaux consacrés à la désambiguïsation des sens des mots des documents textuels en Arabe à base de connaissances, en classant ces travaux selon le type de ressource linguistique utilisé, et en identifiant les défis de chaque type de ce domaine de recherche.

3.2.1. Dictionnaires et thesaurus

Diverses tentatives ont été faites avec les dictionnaires électroniques, dont on essayait d'extraire une information lexicale et sémantique. Cependant, une information rigoureuse n'est pas facile à obtenir, ces dictionnaires présentant deux inconvénients majeurs : ils comportent de grandes incohérences et sont conçus pour être utilisés par des humains, sans tenir compte des besoins logiciels.

Les thesaurus sont aussi exploités comme ressources lexicales pour le traitement automatique de la sémantique. Ils sont plus systématiques que les dictionnaires et fournissent des relations synonymiques entre les mots. Chaque occurrence d'un mot dans une catégorie d'un thesaurus correspond à un de ses sens, chaque catégorie rassemblant des mots ayant approximativement le même sens, notamment pour la constitution du réseau sémantique de [Mast57]. Les méthodes exploitant les thesaurus sont généralement axées sur une information statistique importante, à l'image de [Yaro92], qui établit un modèle statistique basé sur le contexte.

[TlMe06] a testé la validité d'un dictionnaire d'usage afin d'enlever l'ambiguïté du sens d'un mot dans un texte arabe et augmenter le profil de désambiguïsation en utilisant l'algorithme de LESK. Dans la première phase [TlMe06] a représenté du sens par une approche thématique (vecteurs conceptuels), ensuite, il a présenté la démarche suivie pour la construction d'un dictionnaire d'usage. Dans la deuxième phase il a utilisé l'algorithme de LESK qui s'appuie sur la notion du contexte, il permet ainsi d'augmenter le profil de désambiguïsation. L'approche qu'il a effectuée a montré que les variantes d'une méthode de désambiguïsation assez simple, comme l'algorithme de LESK, pourraient produire des résultats comparables à d'autres techniques, plus compliquées ou nécessitant des ressources coûteuses ou difficiles à construire.

[TaYB09] [MeZZ09] ont présenté un moteur de recherche pour la langue arabe qui prend en compte un des niveaux de la sémantique (la terminologie des mots), pour cet objectif, il ont

construit, dans la phase d'indexation, des petits dictionnaires terminologiques pour chaque terme. L'évaluation de leur système se fait par les trois mesures de similarités suivantes : Harman, Croft et Okapi, et les meilleurs résultats sont obtenus par la mesure Okapi. La performance de ce système en termes de précision et de rappel s'est légèrement améliorée. Dans ce travail, [TaYB09] [MeZZ09] supposent que tous les termes des dictionnaires ont la même pertinence. Il conclue qu'il serait intéressant de calculer l'intérêt spécifique pour chaque terme dans son dictionnaire.

3.2.2. Ontologies

Les ontologies sont des ressources linguistiques conçues pour être exploitées par une application logicielle, et qui rassemblent sous la forme de bases de connaissances des informations plus ou moins liées au lexique au niveau morphologique, syntaxique et/ou sémantique. La désambiguïsation sémantique exploite l'ontologie, dont l'information peut se rapprocher tantôt d'un dictionnaire (définitions), tantôt d'un thesaurus (groupes de mots quasi-synonymes, hiérarchie conceptuelle), ou bien d'un réseau sémantique (relations hyponymiques, méronymiques, antonymiques), etc.

Diverses études se sont basées sur différents types d'ontologies pour lever l'ambiguïté dans le texte Arabe.

3.2.2.1. Les travaux basés sur l'ontologie Arabe WordNet (AWN)

[AbBR09] a mis en place un prototype d'un système Question/Réponse (Q/R) Arabe. Ce système est basé sur un module d'expansion sémantique de requêtes utilisant une ontologie lexicale et conceptuelle. L'extension sémantique qu'il a utilisée s'appuie sur l'ontologie Arabe WordNet (AWN) [EBVF06] et la plateforme Amine qui permet le développement de systèmes intelligents [Kabb06].

Dans une autre étude, [AbBR08a] [AbBR08b] ont présenté une évaluation d'un système proposé pour l'expansion sémantique des requêtes basé sur AWN et quatre de ses relations sémantiques. Ils ont utilisé un processus QE (Query Expansion) basé sur : (1) QE par synonymes ; (2) QE par définitions ; (3) QE par sous-types ; (4) QE par sur-types. Deux types d'expériences sont menés : l'évaluation basée sur les mots clés qui utilise un moteur de recherche classique comme passage à un système de recherche d'information et l'évaluation basée sur la structure qui utilise le système JIRS (Java Information Retrieval System), lequel prend en considération la structure de la requête. Dans ce travail, [AbBR08a] [AbBR08b] ont confirmé que le QE sémantique améliore la précision et le Rang Réciproque Moyen (The Mean

Reciprocal Rank MRR). En outre, et dans le cas où il est combiné avec JIRS, cette approche a obtenu une précision autour de 19,51% et 7,85 comme MRR. Cela signifie que lorsqu'on prend en compte la sémantique et la structure de la question, on améliore la probabilité d'obtention de réponses pertinents.

[MoAA10] ont montré une architecture globale pour un moteur de recherche sémantique basé sur une ontologie arabe de vocabulaire limité. Ils ont utilisé uniquement deux exemples pour démontrer l'efficacité de cette architecture, ce qui remet en question le degré d'efficacité de cette évaluation. Ils ont également comparé les résultats de ce moteur de recherche sémantique avec celui d'un moteur de recherche syntaxique (Google), et leurs résultats ont montré que la recherche a été plus précise et réduit l'ambiguïté.

[AEAC13] ont développé une approche dont sa puissance a été prouvée pour la langue anglaise. L'idée est d'exploiter une ressource lexicale (AWN) pour indexer les documents ainsi que la requête de l'utilisateur afin d'améliorer les résultats de recherche. [AEAC13] ont prouvé que les ressources sémantiques améliorent la qualité des systèmes de recherche d'informations par des expériences sur un corpus de moyenne taille de la langue arabe. [AEAC13] ont également remarqué que l'utilisation de la méthode d'indexation sémantique pour représenter les documents et les requêtes ensemble donne de meilleurs résultats que l'utilisation séparée. Ils ont conclu ainsi que la contribution des ontologies dans le système de recherche d'information en langue arabe était très intéressante mais elle nécessite des ressources lexicales complètes qui ne sont pas disponibles à l'heure actuelle.

[HaOL16] de leur côté, ont proposé d'utiliser les deux ressources externes Arabic WordNet (AWN) et WN en employant un système de traduction automatique (terme à terme). De plus, leur travail consistait à choisir le concept le plus proche pour les termes ambigus, basé sur davantage de relations avec des concepts différents dans le même contexte local. Pour évaluer la précision de la méthode proposée, plusieurs expériences ont été menées à l'aide de méthodes de sélection de caractéristiques, Chi-Square et CHIR (techniques d'apprentissage automatique), bayésien naïf et machine à vecteurs de support (SVM). Les résultats obtenus montrent que l'utilisation de la méthode proposée augmente considérablement les performances de leur système de catégorisation des textes en arabe. Ces résultats restent à confirmer sur des corpus plus conséquents.

3.2.2.2. Les travaux basés sur les ontologies spécifiques

[BeAJ10] ont conçu et mis en œuvre un moteur de recherche sémantique en arabe à base d'ontologie, appelé « SemARAB ». L'outil a été construit sur la base d'une similarité sémantique entre des concepts d'ontologie spécifique et une similarité basée sur le contenu de différentes ressources. Cet outils a été destiné au domaine du « commerce électronique », et son évaluation est faite sur cette base. [BeAJ10] ont construit leur ontologie du domaine de « e-commerce » en langue arabe en analysant plus de 100 sites de commerce électronique tels que « souq.com » et « adabwafan.com » afin de déterminer la structure de l'ontologie. [BeAJ10] ont fait une comparaison entre SemARAB et certains moteurs de recherche (Google and Bing), pour montrer la capacité de SemARAB de répondre aux requêtes des utilisateurs et de retourner les résultats les plus pertinents par rapport à ces moteurs. Cependant, SemARAB étant dépendant d'un domaine restreint, il se limite à répondre aux requêtes relatives à une ontologie particulière. Ainsi l'évaluation de ce moteur reste limitée parce qu'elle a été faite sur un échantillon de petite taille (quelques mots).

Dans une autre étude, [ElAA15] ont proposé un nouveau modèle booléen de la recherche d'information sémantique en arabe basé sur l'utilisation des ontologies pour représenter la signification et les relations de chaque mot dont l'index est basé sur son contexte. Les trois ontologies Arabe (Science - Electronics - nature) utilisées ont été construites à l'aide du logiciel Protégé[24] 3.4.3. Ainsi [ElAA15] ont mesuré la précision du modèle proposé et du modèle booléen traditionnel en ayant utilisé des requêtes avec les trois opérateurs booléens (AND, OR, NOT). Les résultats montrent que la nouvelle approche a amélioré la précision mais elle n'a pas amélioré le rappel. En outre, l'évaluation de cette approche reste limitée parce qu'elle a été faite sur un nombre limité de requêtes.

[MRRZ14] ont construit une ontologie à partir des pages Wikipedia et d'autres thésaurus (Arabic Wikipedia Dump, le dictionnaire "Al Raed" et le dictionnaire « Google_WordNet ») afin de l'utiliser dans une approche d'expansion de requête pour améliorer la précision de recherche en la langue arabe. [MRRZ14] se sont concentré sur quatre caractéristiques pour la recherche sémantique : le traitement des subsomptions, le traitement des variantes morphologiques, le traitement de l'appariement de concepts et le traitement des synonymes

[24] Protégé est un système auteur pour la création d'ontologies. Il a été créé à l'université Stanford. Il est gratuit et son code source est développé en Java (voir https://protege.stanford.edu).

(désambiguïsation). [MRRZ14] ont affirmé que les méthodes à base ontologique ont donné des meilleurs résultats que la méthode classique sur la base de mots clés.

3.2.3. Les travaux basés sur l'analyse de corpus

Les méthodes basées sur l'analyse de corpus textuels s'adaptent bien à l'élaboration des modèles statistiques qui reposent sur l'analyse de fréquences rencontrées dans les textes. Cependant, des méthodes linguistiques basées sur des observations et sur la construction de règles à partir de ces observations ont abondamment utilisé les corpus pour obtenir l'information dont elles avaient besoin.

[BEES11] ont proposé une approche pour traiter les termes pertinents du domaine à partir de corpus arabes semi-structurés. En entrée, la structure des documents est exploitée pour organiser les connaissances dans un graphe contextuel, qui est exploité pour extraire les termes pertinents. Ce réseau contient des noms simples et composés gérés par un analyseur morphosyntaxique peu profond. Les expressions nominales sont évaluées en termes du degré de spécificité (« *termhood* » en anglais) et du degré d'unité (« *unithood* » en anglais). Ils ont aussi appliqué une approche qualitative qui pondère les termes en fonction de leurs positions dans la structure du document. En sortie, les connaissances extraites sont organisées comme des dépendances de modélisation de réseau entre termes qui peuvent être exploitées pour déduire des relations sémantiques. L'évaluation de cette approche a été faite sur trois corpus de domaines spécifiques dans le but est de vérifier si ce modèle d'organisation et d'exploitation de la connaissance contextuelle améliorera la précision d'extraction des noms simples et composés. Le résultat démontre empiriquement que ce modèle d'organisation de la connaissance contextuelle basée sur la structure de documents a un impact important sur le processus d'extraction de terminologie. Cependant, la précision de cette approche demeure liée à la qualité du corpus.

Dans un autre travail, [ZoMZ12] [ZZAM12] [Merh09] ont proposé d'utiliser les mesures de Harman, Croft et Okapi avec l'algorithme de Lesk pour développer un système de désambiguïsation de sens de mot arabe. Ils ont utilisé les mesures de RI pour estimer le sens le plus pertinent du mot ambigu. Ainsi, il ont utilisé l'algorithme de Lesk pour identifier le sens adéquat parmi ceux proposés par les mesures de RI. Cette identification est basée sur une comparaison entre les gloses[25] du mot à désambigüiser et ses différents contextes d'utilisation

[25] Explication de quelques mots obscurs d'une langue par d'autres mots plus intelligibles

extraits d'un corpus. [ZoMZ12] [ZZAM12] [Merh09] ont collecté du Web (Wikipédia, le corpus de l'arabe contemporain [AlAt04], le corpus Arabe Coranique, etc.) une grande quantité d'informations, et ils ont formé pour chaque sens possible S_i d'un mot arabe ambigu MA_j, un contexte dont l'ensemble de mots représentent autant que possible du sens S_i. [ZoMZ12] a confirmé que leur étude expérimentale prouve que l'utilisation de l'algorithme de Lesk avec les mesures de Harman, Croft et Okapi permet d'obtenir un taux d'exactitude de 73% (78% pour [ZZAM12]). Cependant, cette étude reste limitée parce qu'il a été faite sur un échantillon de dix mots arabes (cinquante mots pour [ZZAM12]).

3.2.4. Les travaux basés sur des ressources lexicales alternatives

Diverses études se sont basées sur différentes ressources linguistiques et sur l'hybridation entre différentes approches pour lever l'ambiguïté dans le texte arabe.

[ZENM10] a proposé une nouvelle mesure TF-IDF-Okappi qui tient en compte la notion de voisinage sémantique à l'aide d'un calcul de similarité entre termes en combinant le calcul du TF-IDF-Okappi avec une fonction noyau à base radiale afin d'identifier les concepts pertinents qui représentent le mieux un document. En outre, [ZENM10] ont mis au point un dictionnaire sémantique auxiliaire qui est un dictionnaire hiérarchisé contenant un vocabulaire normalisé sur la base de termes génériques et de termes spécifiques à un domaine. Cependant, il ne fournit qu'accessoirement les définitions, les relations entre termes et leur choix l'emportant sur les significations. Les relations communément exprimées dans un tel dictionnaire sont : les relations taxonomiques (de hiérarchie), les relations d'équivalence (synonymie) et les relations d'association (relations de proximité sémantique, proche-de, relié-à, etc.). Dans cette étude, qui ont également utilisé la notion de réseau sémantique comme outils de renforcement du graphe sémantique issu des termes extraits des documents d'apprentissage pour améliorer la qualité et la représentation des connaissances liées à chaque thème de la base documentaire. Cette méthode a permis d'améliorer d'une manière significative les performances de leur système d'indexation. Ces résultats restent à confirmer sur des corpus plus conséquents.

[INHK13] a proposé d'utiliser AWN et une base de connaissance spécifique (Ontology-based Summarization System for Arabic Documents : OSSAD) pour présenter un système de résumé pour les documents arabes. Ces connaissances ont été extraites à partir d'un corpus arabe et ont été représentées par des concepts (/mots-clés) liés au sujet et aux relations lexicales

qui les unissent. Les résultats montrent que le travail de [INHK13] atteint un bon niveau de performance.

Au cours de cet examen des ressources linguistiques utilisées pour gérer le problème de l'ambiguïté sémantique lexicale, nous avons fait plusieurs observations qui doivent nous servir pour le choix d'une ressource adéquate à nos besoins. Permierement, on a remarqué très tôt que l'étude du contexte de la cible constituait la principale information qui permet d'en sélectionner le sens adéquat. Deuxiement, on a constaté l'importance des éléments du contexte syntaxiquement liés à la cible pour effectuer ce choix.

Nous voulons utiliser une méthode de désambiguïsation dans une perspective d'enrichissement et d'expansion de texte, c'est-à-dire que le choix d'un sens correct doit également permettre la sélection d'une information qui s'y rattache. Il s'agit donc d'utiliser une ressource lexicale descriptive qui comporte un maximum de données morphologiques, syntaxiques et, bien entendu, sémantiques.

Dans la section suivante, nous allons présenter les caractéristiques du dictionnaire Arabe « le dictionnaire de la langue arabe contemporaine », lequel est capable de fournir toutes ces informations, et sur lequel nous avons travaillé.

3.3. Le dictionnaire de la langue arabe contemporaine[08مختا]

Comparé à l'Occident, où l'on prend soin des dictionnaires de langue et pour l'élaboration desquels on met tous les moyens à disposition, le monde arabe souffre d'un manque terrible dans ce domaine alors que la langue arabe contemporaine est en developpement continu aussi bien sur le plan de la sémantique que sur le plan des neologisme. Ce qui necessite l'élaboration de dictionnaires de langue qui tiennent compte du progrès scientifique et technologique que connait l'humanité de nos jours.

Les dictionnaires de la langue arabe contemporains, qui ne sont dans leur majorité que des reproductions de ce qu'a été fait, présentent un manque remarquable en matière de néologismes et des mots créés suite au développement scientifique et technologique ; d'où la nécessité à un nouveau recensement des mots de la langue arabe contemporaine avec leurs collocations, contextes et usages dans le but d'élaborer un dictionnaire contemporaine de la langue.

Partant de ce besoin, le Dictionnaire de la langue arabe contemporaine « معجم اللغة العربية المعاصرة » vient combler ce vide dans le but d'integrer les usages moderne des mots dans diffents

pays arabes en evitant toutes les lacunes des dictionnaires précedents qui peuvent se résumer comme suit :

1. Confusion entre ce qui est désuet et ce qui d'usage, et l'absence de beaucoup de mots modernes,
2. L'interdépendance des dictionnaires de langue existants sans examination ni vérification aucunes,
3. L'incapcité de traiter les informations semantiques et morphologiques quant à leurs entrées,
4. L'absence de presque toutes les collocations qui sont très utilisées, ainsi que les expressions contextuelles qui ont acquis de nouveaux sens en plus de leurs sens.

Le Dictionnaire de la langue arabe contemporaine dans son élaboration se base sur un système qui lui est spécifique quant à la presentations des matières, la manière de les présenter et le type d'informations données. Ces dernières comportent le coté morphologique du mot ainsi que son coté sémantique et tous ses usages à travers une étude exhaustive des tous les mots et textes tout en justifiant par des exemples et des expressions contextuelles. Ce dictionnaire accorde aussi une grande imoprtance à la teminologie, comptant dix-mille termes dans differents domaines. Il est également riche en matière de mots courants en se basant sur cosrpus linguistique très riche dépassant un million de mots dans differents contextes. Ce corpus a permis également de recensser toutes les collocations d'uns mot, et surtout si ce dernier est utilisé avec des connecteurs et prépositions; ce qui facilite la connaissance des differents usages d'un mot.

Ce dictionnaire est présenté en deux versions : version papier et version electronique, cette dernière étant très rapides dans la recherche de l'information avec un moteur de recherche très développé qui permet la recherche dans tous les détails du dictionnaire facilitant ainsi à l'utilisateur la recherche de n'importe quel mot ou expression.

3.3.1. Méthodologie du « dictionnaire de la langue arabe contemporaine »

3.3.1.1. Types d'entrées du dictionnaire de la langue arabe contemporaine

Les entrées de ce dictionnaire son classées en cinq types :

- Le verbe : ou l'on ne mentionne pas son type, mais l'on donne les informations morphologiques le concernant ;
- Le nom singulier: suivi de son type [مفرد] ;

- Le nom duel: suivi de son type [مثنّى] ;
- Le nom du pluriel: suivi de son type [جمع];
- les mots fonctionnels: ce sont des mots ayant acquis un nouveau sens loin du sens dénotatif. Ils comprennent les lettres de l'alphabet, toutes les prépositions et les mots interrogatifs, les pronoms relatifs et demonstratifs, les conjonctions de condition, les adverbes, les noms des verbes, les noms des verbes et les verbes figés comme « عسى ».

3.3.1.2. Informations fournies dans le dictionnaire de la langue arabe contemporaine

Le dictionnaire de la langue arabe contemporaine fournit des informations morphologiques et sémantiques sur ses entrées (verbes, noms et mots fonctionnels).

3.3.1.2.1. Informations morphologiques

Elles ne concenent que les verbe et les noms; ce sont des informations morphogiques sur les entrées verbales : le présent, l'impératif (au cas des verbes irrégulies ou des verbes difficle à reconnaitre); le nom d'action (régulier ou irrégulier); le sujet (régulier ou irrégulier); le complémént d'objet (formé à partir d'un verbe transitif direct ou indirect); et le détaché du verbe trilitère pour en preciser le modèle morphologique, exemple « صَبَبْتُ ».

3.3.1.2.2. Informations sémantiques

Elles concenent les verbes, les noms et les mots fonctionnels. Elles sont fournies après l'entrée (avec ses sens terminologique et lexical, et une explication le cas échéant), suivies d'autres exemples (si le sens l'exige), d'un commentaire sur les exemples si c'est necessaire, des expressions contextuelles (si elles existent) avec commentaires (le cas échéant) et l'annotation d'une entrée à une autre si besoin est.

3.3.1.3. Choix de la matière dictionnairique

3.3.1.3.1. Les entrées

3.3.1.3.1.1. Critères du choix des entrées dictionnairiques

1. Les mots courants utilisés ou utilisables au sein de l'intelligentia au temps moderne et les néologismes qui sont des inventions : de la vie moderne tels que (عصرنة، علمانية) , des mots appartenant à la civilisation (سفير، مكوك) , et des termes de la science et de l'art dont l'usage est devenu courant tels que (تلكس، بوصلة). A cela s'ajoutent tous les

mots que le système de la langue arabe accepte et ceux adoptés par les académies et les conférences de la langue arabe.

2. Les noms relevant de la botanique sont pris en détail en raison de la diversité de dénomination d'un pays arabe à un autre.
3. Le choix des entrées dictionnairique est basé sur un balayage automatique de milliers de textes contemporains dans le but de sélectionner des mots nouveaux.

3.3.1.3.1.2. Règles spécifiques pour l'élaboration des entrées du dictionnaire

1. L'entrée du dictionnaire doit être au singlier sauf si l'usage du pluriel est plus courant, par exemple « أساطين » dont le singulier est « أُسطُون ». Il est à signaler que le nombre des entrées classés au pluriel s'élève à 1362. Il est de même pour le duel « المثنى ».
2. Quant aux entées verbales, toutes les prépositions introduisant le premier complément d'objet sont présentées dans ce dictionnaire.
3. Si l'entrée est un verbe transitif par plus d'une préposition, ces prépositions sont classées alphabétiquement. Par exemple le verbe « ضرَبَ » frapper, qui peux être suivi par les prépositions « إلى », « ب », « على », « عن » et « في ».
4. Si un verbe ayant différents sens se présente sous plusieurs rubriques, chacun de ces dernières est classée séparément. Et si plusieurs verbes ayant un sens commun, leurs rubriques sont mentionnées dans une entrée.
5. Si une entrée nominale est voyellisé de plusieurs manières, par exemple « مَثْحَف », « مُثْحَف », l'entrée est écrite de cette manière « مَثْحَف/مُثْحَف », tout en suivant l'ordre alphabétique. Les entrées différemment voyellisées ne sont mentionnées que si leur usage est courant.
6. Le verbe à la forme passive est considéré comme entrée indépendante s'il est corollaire de la structure passive, comme « جُنَّ », « حُمَّ » ou si son usage à la forme passive est courant, comme « صُرِعَ ».
7. Les infinitifs des verbes trilitères et les adjectifs qualificatifs de par leurs irrégularités sont classés dans des entrées indépendantes Des entrées indépendantes sont attribuées aux dérivés réguliers, lorsqu'ils sont mentionnés, et ce dans le but de faciliter la tâche à l'utilisateur du dictionnaire. De même si les dérivés réguliers ont acquis de nouveaux sens comme les terminologies : des entrées indépendantes leur sont attribuées.
8. Les dérivés réguliers avec, le cas échéant, des informations morphologiques irréguliers sont classés séparément. Par exemple « قانِت » dont le pluriel est « قُنَّت » (irrégulier) et « قانِتون » (pluriel irrégulier).

9. Les entrées (verbales ou nominales) sont répétées si les informations morphologiques sont différentes, avec séparation de ces informations et numérotation de l'entrée répétées. Par exemple : « يَمين1 », « يَمين2 », « يَمين3 ».
10. L'entrée doit être indéfinie, avec explication en termes indéfinis. Et Si le vocable est corollaire de la définition, ou bien connu sous sa forme définie, il est mentionné dans l'entrée sous sa forme indéfinie, puis sous sa forme définie dans une autre case.
11. Pour la correction linguistique et l'approbation des mots et leurs usages, les décisions de l'académie égyptienne de la langue arabe sont prises en considération.

3.3.1.3.2. L'entrée à travers un exemple

C'est un type ou l'usage de l'entrée n'est utilisé qu'à travers un exemple dans le but de déterminer le sens. Pour les noms on prend l'exemple « يوم التغابن » : « يوم القيامة », et pour les verbes, dont le sens, la transitivité et autres caractéristiques ne sont déterminées qu'à travers un exemple de phrase, on prend l'exemple de « زأرَ الأسدُ » « le lion rugissait » ou le faculté de rugir n'est attribuée qu'au lion dans ce cas.

Lors de l'élaboration des entrées du dictionnaire, il a été pris en considération de plusieurs spécificités, à savoir :

1. Des structures traditionnelles au sens coranique tel que : « تابوت العهد » (l'arche d'alliance).
2. Si l'entrée verbale directement transitive et indirectement transitive a le même sens, elle est mentionnée suivant l'exemple : « دأَب فلانٌ الشَّيءَ/ دأب فلانٌ على الشَّيءِ », et ce afin d'éviter la répétition.
3. L'entrée nominale composée est classée sous le mot le plus remarquable de son expression. S'il en existe plusieurs, l'entrée est répétée suivant le nombre des mots remarquables.

3.3.1.3.3. Les exemples supplémentaires

1. Les extra-exemples couvrent tous les contextes du mot, ce qui explique l'abondance des exemples : des exemples du Coran, des lectures, des Hadiths, des poèmes, des proverbes...etc. le nombre de ces exemples s'élève à 43385.
2. La priorité est donnée aux exemples du Coran pour son éloquence et aux collocations pour leur usage répandu. Par exemple : « أخذ حذره », « فتح الجلسة », « فتح المظاريف », « فحص المريض ».

3. Les mauvais exemples et les exemples élaborés ont été évités, tout en choisissant des exemples ayant un sens culturel et moral.
4. Les exemples couramment utilisés sont mis à jour, exemple on ne dit pas seulement « ترجل عن الدابة », mais aussi « ترجل عن السيارة ».
5. Pour la citation, les exemples choisis sont ceux les plus courts et les plus explicites.
6. Les commentaires sur les exemples sont en courtes expressions.
7. Le « Hadith » englobe aussi les paroles des « Sahaba » et « Tabi'ne » (Compagnons et Suiveurs) suivant les méthodes des dictionnaristes adoptées.
8. Peu d'exemple de poésies ont été évoqués en prenant en considération la clarté et l'authenticité.

3.3.1.3.4. Les expressions contextuelles, collocations et structures

Pour les expressions contextuelles, il a été pris en considération de :

1. Qu'elles soient celles du temps moderne et courantes.
2. Qu'elles soient classées selon le vocable le plus remarquable.
3. Qu'elles soient mises devant le sens le plus proche.
4. Commenter les expressions qui nécessitent un commentaire pour être comprises.

3.3.1.3.5. Les sens

1. Peu d'intérêt est accordé aux informations encyclopédiques, plus précisément les informations historiques.
2. Les mots fonctionnels et les terminologies sont expliqués.
3. Le type du dérivé et le genre du mot est indiqué si c'est utile.
4. L'explication des analogies et des groupes de mots clés est unifiée.
5. La répétition de relation entre différents sens a été évitée.

3.3.1.4. Méthodes d'interprétation suivies dans le dictionnaire

Les méthodes suivies sont :

1. Explication par synonyme ;
2. Explication par antonyme ;
3. Explication par définition ;
4. Explication par exemplification réelle.

Et pour se faire, il a été pris en considératon de :

1. Éviter d'utiliser des termes techniques ;
2. Vérifier la précision lors de la collecte des sens proches et faire la différence entre les sens distincts ;
3. Ne pas utiliser des mots étranges dans l'explication ;
4. Ne pas utiliser les définitions générales ;
5. Si le sens est terminologique, une abréviation de la science sous laquelle le terme se classe est mise entre des crochets au début du sens.

3.3.1.5. Le système de référence

Le système de référence d'une entrée à une autre a été suivi dans plusieurs cas, notamment:

1. Lorsqu' il existe une relation entre deux entrées différentes ayant un même sens, comme « كونفدرالية » et « تحالف » ;
2. Ainsi que lorsqu'il existe plusieurs formes du mot, telles que : « آزورت » et « أزوت », « أوكسيد » et « أكسيد », « موسيقى » et « موسيقا » où la signification est placée sous l'entrée qui est d'abord placée dans l'ordre, avec l'utilisation du système de référence dans la deuxième entrée ;
3. Lorsque l'entrée est placée sous plusieurs racines, en particulier dans les mots fonctionnels, elle est placée sous les lettres telles quelles et insérée sous sa racine trilitère possible, sans donner ni sens ni information, par exemple : l'entrée « هَلُمَّ » présentée sous la racine « هـ ل م », « هـ ل مـ » ;
4. Quand le mot contient plus qu'une racine arabe correcte, où les informations sont fixées dans chaque endroit avec un sens unifié dans chacun en suivant un système de référence. Par exemple l'entrée « ذُرِّيَّة », dont la racine est « ذ ر أ », sa référence est faite ainsi : voir « ذ ر ر – ذُرِّيَّة » ، « ذ ر و – ذُرِّيَّة ». Ce système de référence est appliqué sur les deux racines : « ذ ر ر » et « ذ ر و » ;
5. Quand il y a deux mots ayant le même sens et la même racine, la définition du sens de chacun d'eux est donnée précédé de l'autre mot, exemple : « قِيثار » et « قِيثارة » dont le sens est fixé « آلة طرب ذات سِتَّة أوتار », et le mot « قيثارة » par « قِيثار، آلة طرب ذات سِتَّة أوتار »;
6. Lorsque l'origine de l'entrée est doutée, telle que l'entrée « ميناء » dans son ordre alphabétique, la référence à sa position d'origine est faite sous la racine « و ن ي » ;

7. Le système de référence a été utilisé pour les mots arabisés et d'origine non arabe pouvant éventuellement inclure des caractères alphabétiques supplémentaires : le mot « استبرق » est placé sous la racine « ب ر ق » et une référence est faite à « استبرق », etc.

3.3.1.6. Règles générales

1. Les racines sont écrites en lettres séparées ;
2. Les racines avec les lettres « و » et « ي » sont fixée ainsi : « ف ت و/ ف ت ي », car nulle racine n'est pondérée par les dictionnaires par rapport à l'autre racine ;
3. Le verbe irrégulier sourd « المضعَّف » sous sa double forme (par séparation ou confusion des deux dernières lettres identiques) à l'impératif est donné, comme le verbe « افْكُكْ/ فُكَّ ». Il est de même pour le verbe Hamzé « المهموز », comme « اسْأَلْ/ سَلْ » et « اوْخُذْ/ خُذْ » ; et également le verbe faible « المعتل » ayant plusieurs formes, comme « وعى : عَ / عَهْ » ou la lettre « هـ » et fixée ou supprimé ;
4. Pour la facilitation de l'usage du dictionnaire, des entrées indépendantes ont été réservées pour le mots fonctionnels qui change selon le nombre, le type ou bien la déclinaison syntaxique, comme « الَّذي، اللّذان، الَّذين، اللذَيْن، الَّتي، اللّتانِ، اللّتَيْن، اللاتي، الّلائي، اللَّواتي » ;
5. La qualification du complément d'objet formé à partir d'un verbe transitif est distinguée de celle formée à partir du verbe intransitif par le mot "للمتعدي" mis entre parenthèses ;
6. Si le nom a plus d'un sens dont un est un sens issu d'un infinitif, le pluriel est distingué par l'expression « لغير المصدر » mise entre parenthèses après mention du pluriel, pour affirmer la non-validité du pluriel dans ce cas sauf pour donner le sens du nombre de la multiplicité des types ;
7. Le pluriel régulier n'est mentionné que lorsque le mot a plus d'un type de pluriel (comme : « صمام » dont le pluriel peut être « صمامات » ou « أصمَّة ») ou si le mot est d'origine linguistique étrangère (comme : « بَبَّغاء ») ;
8. L'adjectif qualificatif formé sur le modèle « فعلان » dont le féminin est « فَعْلى/ فعلانة » est indexé. Par exemple l'entrée : « عطشانُ/ عطشانٌ » féminin « عطشى / عطشانة » ;
9. Les attributs (les adjectifs de relations) exceptionnels ont été mentionnés, par exemple l'adjectif de relation formé à partir du pluriel (car sa formation à partir du singulier étant impossible) ;
10. Pour la transcription des versés coraniques, il a été pris en considération de l'édition de « مصحف المدينة النبوية », selon la lecture de HAfs d'après Assim ;

11. Le système orthographique est unifié dans tout le dictionnaire.

3.3.1.7. Des statistiques

Les statistiques des différents types d'entrées dans le dictionnaire de la langue arabe contemporaine reflètent la richesse linguistique de cette ressource lexicale (Tableau 3.1).

	L'Entrée	*nombre*
01	Racine	5778
02	Toute (Noms, Verbes, mots fonctionnels)	32300
03	Verbes	10475
05	Noms	21457
06	Singulier	20070
07	Duel	24
08	Pluriel	1362
09	Mots fonctionnels	368
10	Les exemples	29118
11	Les sens	63019
12	Les exemples supplémentaires	43384
13	Les expressions contextuelles	17883
14	Les informations morphologiques	59601
15	Les relations morphologiques dans la sémantique	16012
16	Les concepts	10064

Tableau 3.1: Des statistiques sur le dictionnaire de langue arabe contemporaine

3.4. Conclusion

Le but de ce chapitre était de présenter les travaux sur les approches de la recherche d'information sémantique dans les documents textuels arabes, et plus particulièrement les approches basées sur les ressources lexicales pour la désambiguïsation sémantique des mots.

Dans ce chapitre nous avons présenté les études les plus intéressantes et les plus récentes sur l'utilisation de l'approche à base de connaissance en langue arabe, et nous avons défini et clarifié les caractéristiques du dictionnaire « de la langue arabe contemporaine », lequel nous allons utiliser dans notre étude.

Dans le chapitre suivant, nous allons représenter nos contributions dans le domaine recherche d'information sémantique dans les documents textuels en arabe, et nous

commencerons par la conception et l'implémentation d'un système de RI, en mettant le point sur les approches de lemmatisation des mots et leurs rôles dans la performance des systèmes de RI.

Partie 2 :

Contribution à la proposition d'un modèle de RI sémantique en arabe

Chapitre 4 :

Une méthode de lemmatisation hybride du texte arabe pour un système de recherche d'information sémantique robuste

4. UNE METHODE DE LEMMATISATION HYBRIDE DU TEXTE ARABE POUR UN SYSTEME DE RECHERCHE D'INFORMATION SEMANTIQUE ROBUSTE

4.1. Introduction

L'arabe est l'une des six langues officielles des Nations Unies et la langue maternelle de plus de 400 millions de personnes[26], ce qui représente environ 5.6 % de la population du monde entier. Récemment, en raison du nombre croissant d'internautes dans le monde arabe, la recherche d'information (Information Retrieval : IR) est devenue un outil essentiel pour toutes les tâches de recherche sur le Web. En juin 2017, le nombre d'utilisateurs arabes d'internet s'élevait à environ 185,000,000 millions, ce qui représente environ 43.8 % de la population du monde arabe et environ 4.8 de la population du monde entier. Il existe actuellement peu de moteurs de recherche en arabe par rapport à d'autres langues, malgré les efforts considérables déployés pour répondre aux besoins du nombre croissant d'internautes arabes. De plus, l'arabe est une langue très flexionnelle et possède une structure morphologique complexe [KhGa99] [LaCo01b] [DaOa03] [ChGe02], ce qui fait que la recherche d'informations sur des textes arabes nécessite la forme de base du mot (racine ou lemme) [WiGa07]. Par conséquent, le processus de lemmatisation est nécessaire.

La recherche d'informations dans un texte arabe est devenue de plus en plus importante. Ce domaine de recherche a considérablement progressé au cours des dernières décennies, comme il est la principale motivation pour l'intérêt d'étudier le traitement du langage naturel.

L'arabe a une morphologie très riche et complexe. Son origine est très différente des langues européennes. Il comprend 28 lettres et écrit de manière cursive de droite à gauche. La représentation morphologique de l'arabe est assez complexe en raison de la variation morphologique et du phénomène d'agglutination.

En plus de la morphologie complexe de l'arabe écrit, les voyelles (diacritiques) sont omises, d'où la tendance des mots à avoir un niveau d'ambiguïté plus élevé, ainsi que le problème du pluriel des noms irréguliers. Dans ce cas, un nom au pluriel prend une autre forme morphologique différente de sa forme initiale au singulier.

[26] https://www.internetworldstats.com/stats19.htm

En outre, la lemmatisation des mots arabes dans les systèmes de recherche d'information sémantiques est un processus sensible, parce qu'une indexation sémantique efficace dépend de la désambiguïsation des sens des mots qui doivent avoir des lemmes corrects.

Pour résoudre ces problèmes et bien d'autres, nous nous basons sur les algorithmes de lemmatisation pour regrouper les mots en fonction de la similarité sémantique. Il existe plusieurs types de ces algorithmes. Les deux lemmatiseurs les plus efficaces en arabe sont le lemmatiseur léger de Larkey [LaCo01b] [LaBC02] et le lemmatiseur *d'extraction des racines* de Khoja [KhGa99].

Nous développons un système de recherche d'informations dédié à la langue arabe basé sur une méthode hybride en phase de lemmatisation combinant trois techniques connues : la suppression d'affixes proposée par Kadri [KaNi06a], les dictionnaires [AlEv94] et l'analyse morphologique [Bees98] [Ahme00] [MoMo02].

Dans ce chapitre nous présentons l'architecture de notre système de recherche d'information dans un texte arabe, avant d'ajouter la couche sémantique, et fournissons une analyse complète sur un certain nombre de niveaux liés à la recherche d'informations, en particulier : (I) une étude des différentes méthodes de lemmatisation en arabe, (II) des applications de certaines méthodes de lemmatisation et (III) d'évaluation de la performance de notre contribution qui est une méthode de lemmatisation hybride pour la langue arabe.

4.2. 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, le corpus TREC incluant des documents, des requêtes et des jugements de pertinence est le plus grand corpus en arabe actuellement disponible. Il 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 [Kadr08]. Et comme nous ne disposons pas de ce corpus ou d'un autre qui soit professionnel 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 on a trouvé le site web « Al-Khat Alakhdar ». Ce dernier, spécialisé dans le domaine de l'environnement, est restreint 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 » auquel nous avons intégré quelques requêtes et jugements de pertinences fait l'objet d'une importante production langagière en arabe. Il contient 694 articles couvrant une période bien déterminée et représentant un volume de 15 MOctets.

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 4.1) et un autre de ces requêtes est présenté dans la (Figure 4.2).

<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/>

Figure 4.1: Exemple d'un document du corpus «Al-Khat Alakhdar» [Dile11]

<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/>

Figure 4.2: Exemple d'une requête du corpus «Al-Khat Alakhdar» [Dile11]

Le Tableau 4.1 présente quelques caractéristiques du corpus « 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 4.1 Caractéristiques du corpus «Al-Khat Alakhdar» [Dile11]

4.3. Architecture du système RI

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

4.3.1. Indexation

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.

Formellement, si on suppose les notations suivantes :

- $\mathcal{C}$: un corpus ou un ensemble de documents $\{d_1, d_2, \dots, d_n\}$
- n : le nombre de documents du corpus.
- d_i : un document ou une séquence de termes t, notée $\langle t_1, t_2, \dots, t_i \rangle$.
- l_i: la longueur du document d_i
- $\mathcal{T}$: le dictionnaire ou l'ensemble des termes distincts du corpus $\mathcal{C}$.
- $idx_j = (o_1, o_2, \dots, o_n)$ est 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.
- $pos_{jk} = (p_1, p_2, \dots, p_o)$ est 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.

Alors, le processus d'indexation est le calcul de idx_j et de pos_{jk} pour tous les termes t_j du dictionnaire $\mathcal{T}$ du corpus $\mathcal{C}$ [FaGu06].

L'indexation est définie par AFNOR[27] 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 [Asso93].

4.3.2. 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.

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 spécifie 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 et ordonnancer le résultat ;
- Rechercher les noms des documents du résultat.

4.4. L'implémentation du SRI dans le texte arabe « OIRDA »[28]

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

---

[27] **AFNOR** : est un groupe français issu de la fusion des associations : Association française de normalisation (Afnor) et Association française pour l'assurance de la qualité (Afaq) et qui comprend trois filiales commerciales autour de l'association Afnor. AFNOR conçoit et déploie des solutions fondées sur les normes volontaires, partout dans le monde. Il est au service de l'intérêt général dans sa mission de normalisation et exerce dans le domaine concurrentiel des activités de formation, de veille et d'information professionnelle et technique, d'évaluation et de certification (https://www.afnor.org/).

[28] OIRDA : **O**util d'un **I**ndexation et de **R**echerche dans les **D**ocuments textuels **A**rabe

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

- Unifier l'encodage de texte soit pour le corpus, soit pour les requêtes ;
- Normaliser le corpus de textes et les requêtes ;
- Découper ou segmenter le texte d'entrée en séquences d'unités lexicales (mots) ;
- Éliminer les mots vides (Stop Words) ;
- Déterminer pour chaque mot ses caractéristiques morphologiques ;
- Lemmatiser les mots résultants, en supprimant les préfixes et les suffixes sur la base des 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 4.3).

4.4.1. 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 sera transformé en format Unicode dans notre cas.

4.4.2. Normalisation

Comme nous l'avons déjà précisé dans le premier chapitre : la manipulation des variations du texte qui peuvent être représentées en arabe, nous mènent à exécuter plusieurs genres de normalisation sur le texte de corpus (documents et requêtes).

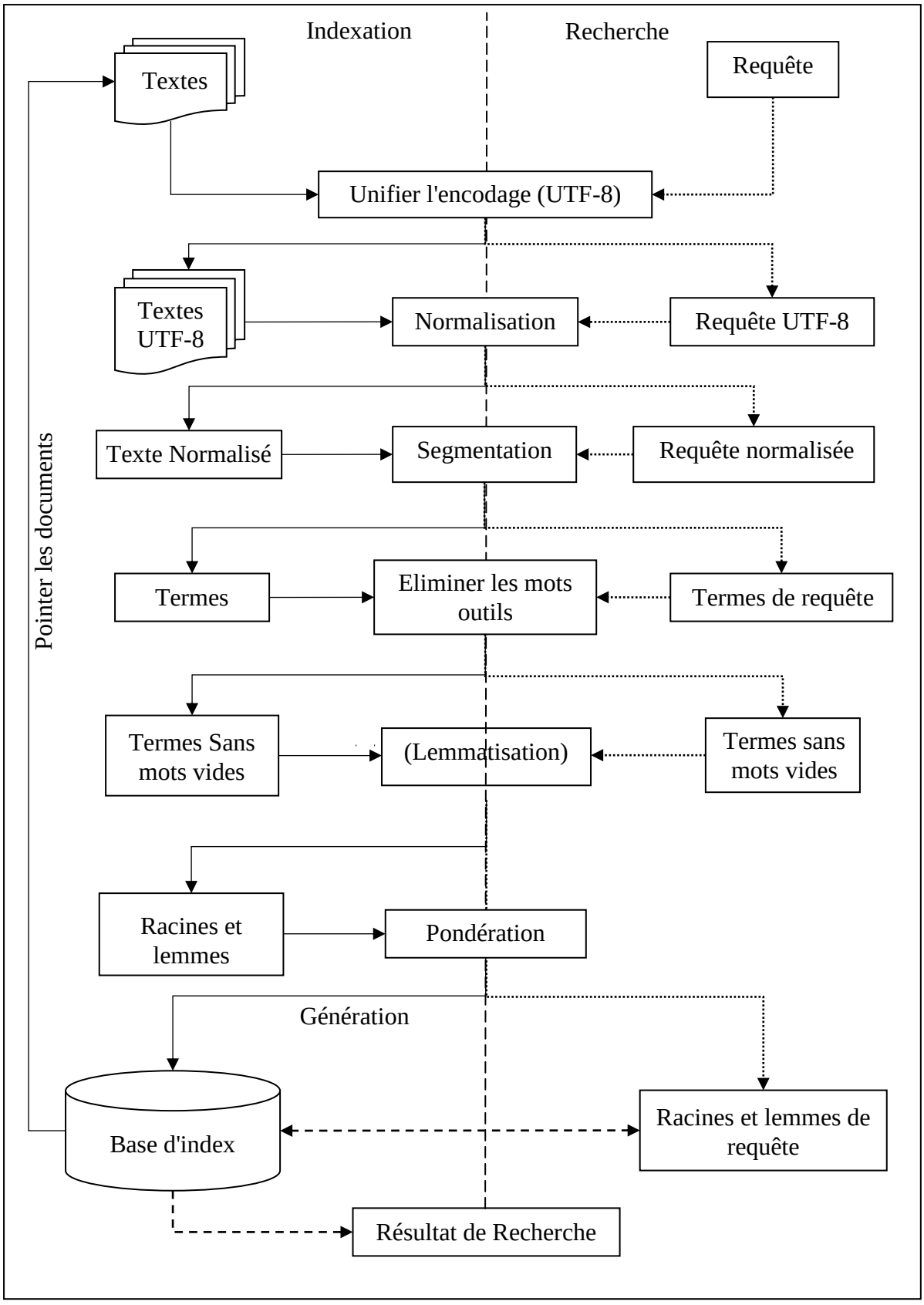

Figure 4.3: Architecture d'un système de RI pour les textes en langue arabe [Dile11]

4.4.3. Segmentation

La segmentation est une étape nécessaire et signifiante dans le traitement du langage naturel. 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 les 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 4.4).

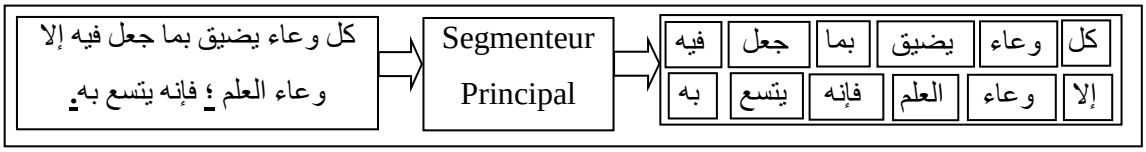

Figure 4.4: Exemple d'un Segmenteur [Dile11]

4.4.4. 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 4.2: Un aperçu sur les mots vides [Dile11]

Il est intéressant de souligner que 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.

4.4.5. Lemmatisation

Comme nous avons vu, le traitement morphologique est le cœur de la recherche d'information pour les textes en arabe, plus précisément c'est la lemmatisation qui joue un rôle important. Alors nous appliquons 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.

4.4.5.1. 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 proposée par Kadri [Kadr08], 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 lemmes, par exemple (لنموه, pour sa croissance) si on retire le suffixe d'abord (وه), alors on va perdre le lemme correct (Figure 4.5).

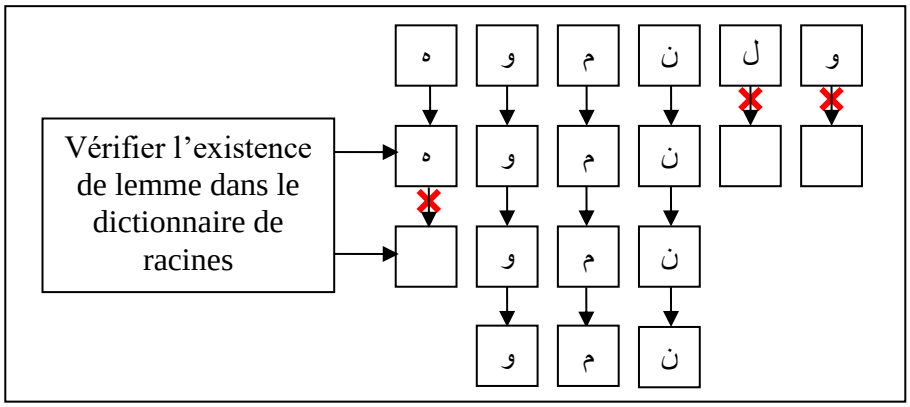

Figure 4.5: Exemple sur la méthode PS-M [Dile11]

4.4.5.2. 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 (ال), alors on va perdre le lemme correct (Figure 4.6).

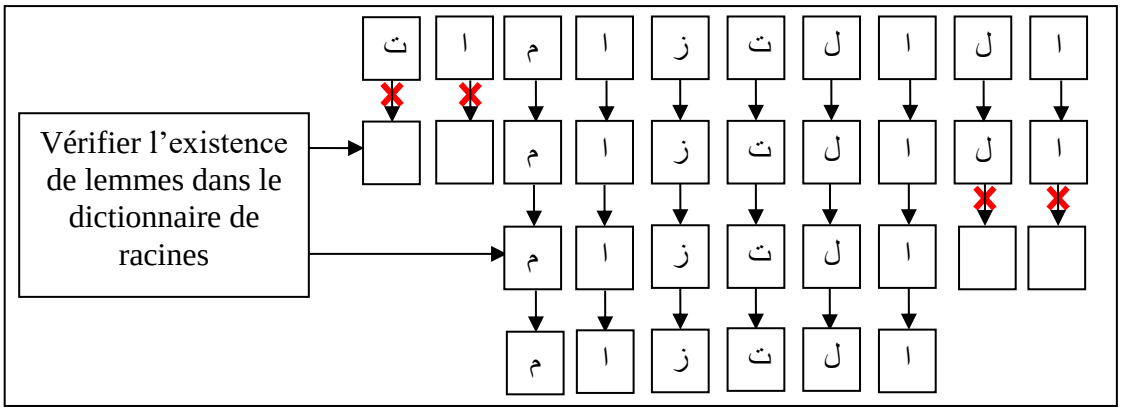

Figure 4.6: Exemple sur la méthode SP-M [Dile11]

4.4.5.3. La méthode PS+M

Après avoir retiré tous les préfixes puis les suffixes du mot fléchi, nous avons comparé celui-ci avec tous les modèles disponibles ; c'est pourquoi, nous avons nommé cette méthode PS+M (Préfixe Suffixe Avec Modèle). Par contre, 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 [AlEv98]. Par exemple nous n'avons conservé aucun de ces modèles : «استفعل », «مستفعل » parce que les deux préfixes « است », « مست » sont existants. 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 ». Ensuite, nous avons retiré ces lettres et extrait la racine « حسب » (Figure 4.7).

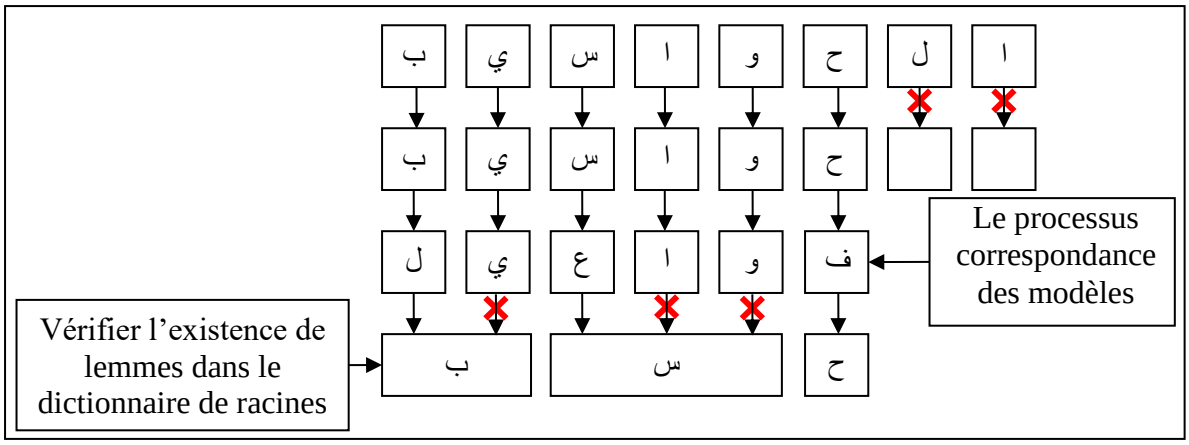

Figure 4.7: Exemple sur la méthode PS+M [Dile11]

4.4.5.4. 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.

4.4.5.5. La méthode HY

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.

L'algorithme global de cette méthode est donné comme suit :

```
Program Arabic_Stemming
  While there is a prefix do
     /*The technical of dictionaries*/
     Check the existence of the word in the dictionary
   If it exists then
    Add to index
    Exit program
   Else
    While there are suffixes do
     While there are models do
     /* The morphological analysis technique */
     Compare the word with the model
      If there is a model then
       Extract the root
       /*The technical of dictionaries*/
     Check the existence of the word in the dictionary
       If it exist then
         Add to index
         Exit program
       End If
      End If
     End While
     /* The affix-removal dictionaries */
     Remove the suffix
    End While
   End Else
   /* The affix-removal dictionaries */
   Remove the prefix
  End while
  Write ("The word is wrong")
End Program
```

Figure 4.8: L'algorithme global de la méthode de lemmatisation hybride

4.4.6. 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 [SaMc86]. 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 [JoWR00] :

$$Poids_i = tf_i \cdot idf_i \qquad \text{(Eq.4.1)}$$

$$idf_i = \log(N/N_i) + 1 \qquad \text{(Eq.4.2)}$$

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.

4.4.7. 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 [ZoMR98]. 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 4.9: Le fichier inverse correspondant à un texte simple [Dile11]).

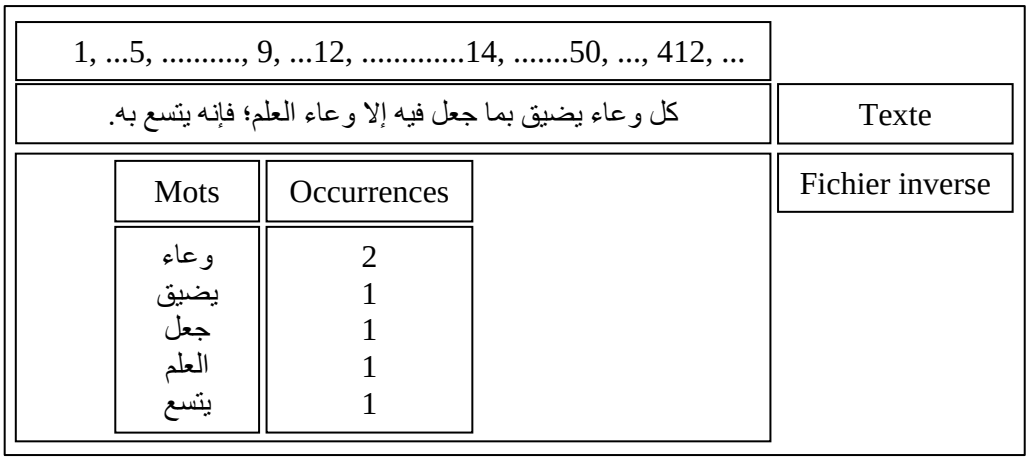

Figure 4.9: Le fichier inverse correspondant à un texte simple [Dile11]

4.4.8. 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.

4.4.8.1. L'appariement document-requête

Avant de décrire le module d'appariement document-requête, il faut rappeler 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.

4.5. 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 [Kadr08].

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 4.3 présente un exemple montrant les résultats des expériences associés à une requête (حرائق النفط).

Nous comparons tout d'abord les deux méthodes de lemmatisation (PS-M et SP-M) que nous avons proposées.

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

La Figure 4.10 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 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.

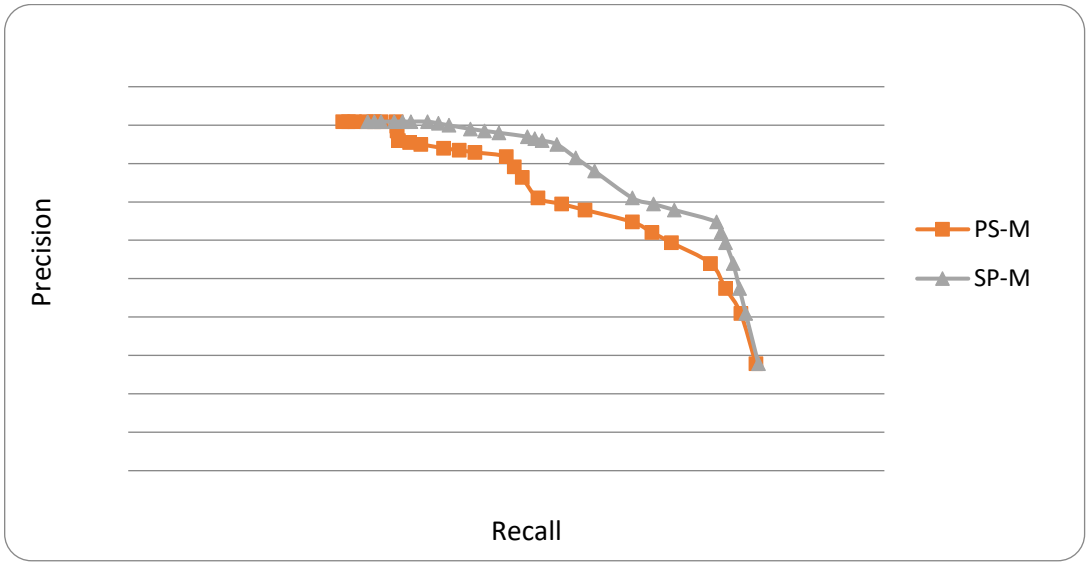

Figure 4.10: Les courbes rappel-précision des deux méthodes de lemmatisation PS-M et SP-M

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 choisir 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 4.11 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.

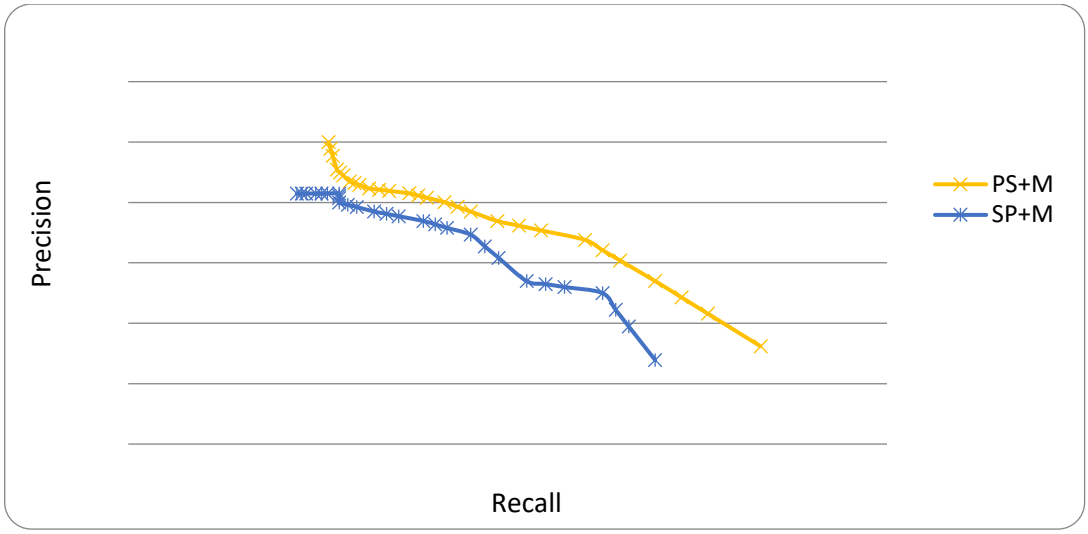

Figure 4.11: Les courbes rappel-précision des deux méthodes de lemmatisation PS+M et SP+M

La Figure 4.12 dresse une comparaison entre les méthodes de lemmatisation sans modèle et avec modèle sur notre corpus en fonction de leurs courbes rappel-précision.

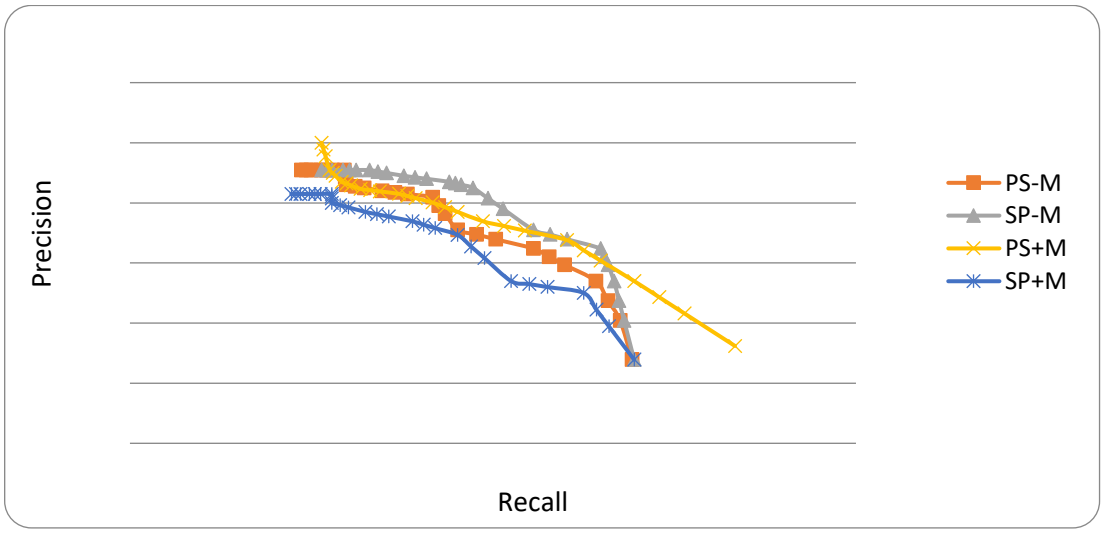

Figure 4.12: Les courbes rappel-précision des méthodes de lemmatisation PS-M, SP-M, PS+M et SP+M

Sur notre corpus, 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 4.12 ; 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 4.12 montre que la méthode de lemmatisation PS+M obtient les meilleurs scores quand le rappel est inférieur à 10 %.

Pour cette raison, nous avons proposé une nouvelle méthode de lemmatisation hybride (HY), qui combine toutes les méthodes mentionnées précédemment pour améliorer la performance globale du processus de lemmatisation. La Figure 4.13 dresse une comparaison entre les cinq méthodes de lemmatisation sur notre collection en fonction de leurs courbes rappel-précision.

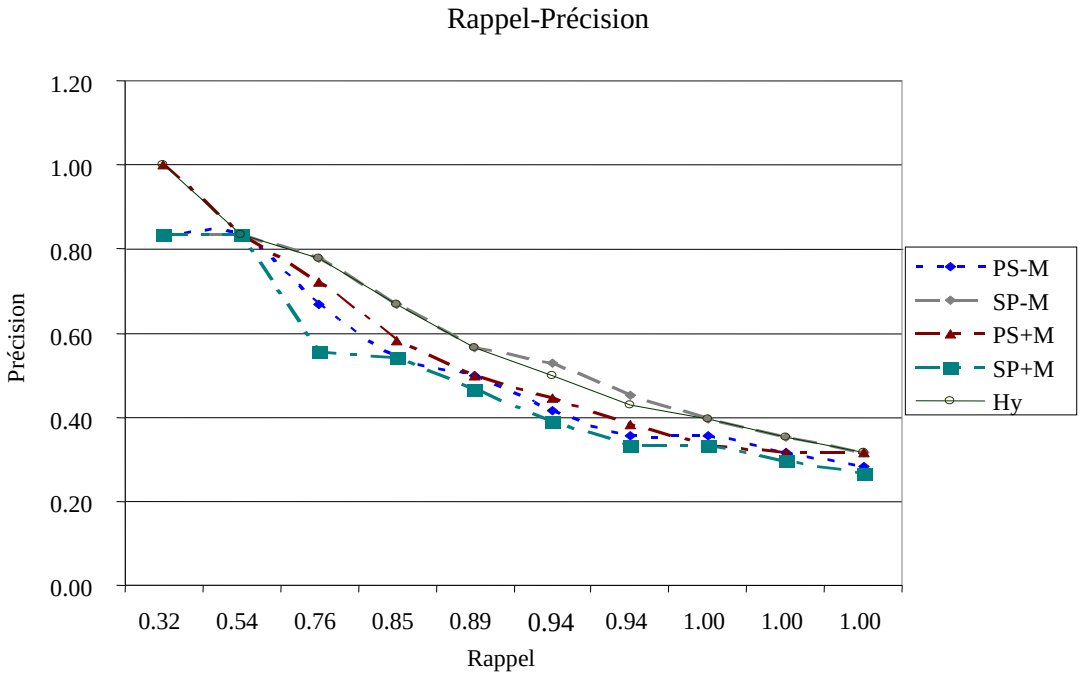

Figure 4.13: Les courbes rappel-précision des cinq méthodes de lemmatisation

Ces résultats montrent que la méthode de lemmatisation HY sont plus efficaces que les autres méthodes. On peut observer ce comportement dans la Figure 4.13 : le courbe de lemmatisation HY représentant 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 contre 57%, 48%, 51%, et 54% pour les méthodes SP-M, SP+M, PS-M, et PS+M respectivement.

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.

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.

4.6. Conclusion

L'arabe est 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 arabes, 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 appliqué 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 laquelle 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 d'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, ces 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.

Le chapitre suivant de la thèse présente l'impact de l'indexation en ligne sur l'amélioration des systèmes de recherche d'information sémantique en arabe.

Chapitre 5 :

L’impact de l'indexation en ligne sur l'amélioration des systèmes de recherche d’information sémantique en arabe.

5. L'IMPACT DE L'INDEXATION EN LIGNE[29] SUR L'AMELIORATION DES SYSTEMES DE RECHERCHE D'INFORMATION SEMANTIQUE EN ARABE.

5.1. Introduction

La grande disponibilité des informations rendait particulièrement difficile l'obtention et la recherche d'informations pertinentes et utiles pour les utilisateurs. Dans ce contexte, les systèmes de recherche d'informations sont apparus comme un outil permettant de résoudre ce problème, ces systèmes comprend deux étapes : les étapes « indexation » et « recherche ». Dans la première étape, les descripteurs sont extraits de documents et préparés pour faciliter et accélérer le processus de recherche dans la deuxième étape. En général, l'étape d'indexation comprend trois types. Premièrement, l'indexation manuelle, dans laquelle le processus de sélection des descripteurs est effectué par un expert humain. Deuxièmement, l'indexation automatique où les descripteurs sont automatiquement extraits des documents et, enfin, l'indexation semi-automatique (ou l'indexation supervisée). Ce dernier fournit une assistance automatisée à l'expert.

Actuellement, les systèmes de RI bénéficient des processus d'indexation, dont la plupart restent sous-performant pour l'extraction de descripteurs précis contribuant à l'amélioration de la qualité de ces systèmes, notamment l'extraction de la sémantique de ces descripteurs. Cela reste une tâche difficile d'indexation automatique qui nécessite souvent une intervention humaine pour choisir les descripteurs appropriés. Cela est dû à plusieurs raisons, notamment l'ambiguïté du langage, le pouvoir du langage de transférer les pensées d'un esprit à un autre et la nature dynamique du langage.

Bien que la littérature se compose de nombreuses études concernant diverses langues naturelles, il y a relativement moins d'études sur la langue arabe, où les caractéristiques grammaticales et morphologiques complexes de cette langue rendent la tâche du traitement automatique encore plus ardue. Ainsi, dans ce chapitre nous présentons un nouveau type d'indexation pour contribuer à l'amélioration de la qualité des systèmes de RI. La méthode d'indexation proposée appartient à la catégorie d'indexation semi-automatique et se compose de deux types. Le premier type effectue une indexation en ligne où un document est l'unité

[29] L'indexation en ligne fait référence au processus d'indexation qui commence directement après la fin de la rédaction de chaque document. Dans ce cas, les textes sont capturés et stockés via différents outils de traitement de texte. Tandis que dans le cas d'indexation hors-ligne, le processus d'indexation est effectué sur la collecte de documents textuels disponibles dans différents corpus.

d'indexation. Ce type d'indexation fait référence au processus d'indexation qui commence directement après la fin de l'écriture de chaque unité, ce qui permet d'aider l'expert humaine (auteur du texte) à sélectionner les descripteurs arabes appropriés pour améliorer les résultats de la recherche. La sortie de ce processus donne lieu à un index partiel. Le second type - sous cette méthode - est une indexation hors ligne (offline), qui fait référence au processus d'indexation basé sur la collecte de documents textuels disponibles à partir de différents corpus. La sortie de ce processus conduit à un « *index général* ».

Nous illustrons également la mise en œuvre et les performances de cette nouvelle méthode d'indexation à l'aide d'un éditeur de texte arabe développé et conçu pour avoir un système d'indexation semi-automatique en ligne et un outil de recherche d'informations contenant un système d'indexation automatique hors-ligne. Nous illustrons également le processus de construction d'une nouvelle forme de corpus arabe permettant de mener les expériences nécessaires.

Ainsi, cette étude contribue à deux domaines clés de la littérature. Premièrement, elle offre des applications de certains outils, tels que « SIRAT[30] » et « OIRDA[31] », qui ont été développés pour montrer à quel point l'intégration de l'indexeur semi-automatique en ligne dans les éditeurs de texte permet d'améliorer la précision et l'indexation sémantique des systèmes de RI. Deuxièmement, l'étude est menée sur des textes arabes, ce qui contribue à l'enrichissement et au développement d'outils de traitement de la langue arabe.

Dans ce chapitre, nous allons offrir un compte rendu des principaux développements et avancées récentes de l'indexation des documents textuels en arabe, et nous allons identifier les principales caractéristiques de la langue Arabe, et illustrer notre système d'indexation semi-automatique, les applications mises en œuvre et les résultats des expériences réalisées.

5.2. L'indexation des documents textuels et l'extraction de mots-clés arabes

Nous commençons par une brève présentation des principaux travaux consacrés à l'indexation des documents textuels en arabe et une identification des défis de ce domaine de recherche, en classant ces travaux selon l'approche la plus utilisée. Nous présentons ensuite des

[30] L'éditeur de texte Arabe SIRAT (**S**emantic **I**nformation **R**etrieval for **A**rabic **T**exts) est une application que nous avons développée pour mener des expériences sur le domaine de recherche d'information sémantique dans le texte Arabe.

[31] OIRDA (**O**util d'**I**ndexation et de **R**echerche dans les **D**ocuments **A**rabes) : C'est un programme d'indexation et de recherche de textes Arabe que nous avons développé en Java.

travaux liés à l'extraction automatique de mots-clés en arabe, qui contribuent à améliorer la qualité des systèmes d'indexation.

5.2.1. L'indexation des documents textuels en arabe

Comme nous avons vu (chapitre 4), diverses études ont proposé de différentes méthodes d'indexation des documents en arabe. Ces études proposent diverses techniques d'indexation automatique selon les approches suivantes : linguistique, statistique, sémantique et hybride. Cependant, à notre connaissance, toutes ces études étaient axées sur l'indexation manuelle et automatique. Cela nous a empêché de comparer les méthodes existantes à celles proposées dans cette étude.

5.2.1.1. L'approche linguistique

L'approche linguistique consiste en une analyse morphologique et syntaxique du document basé sur les règles grammaticales et les relations entre les différentes unités textuelles. Les méthodes de cette approche sont largement utilisées dans le traitement de la langue arabe en raison de la fiabilité des algorithmes de reconnaissance syntaxique et sémantique. [BeST07] ont proposé des systèmes d'extraction de connaissances, basés sur une analyse linguistique approfondie et utilisant une ontologie de domaine pour extraire le contenu sémantique, ont donné des résultats prometteurs, mais révèlent d'autres problèmes nécessitant une enquête approfondie.

[MHDH08] ont proposé une méthode basée principalement sur l'analyse morphologique et sur une technique d'attribution de poids aux mots. L'analyse morphologique utilise un certain nombre de règles grammaticales pour extraire les mots candidats d'index. La technique d'attribution de poids calcule les poids de ces mots par rapport au document conteneur. Les pondérations sont basées sur la dispersion des mots dans un document et pas seulement sur leur taux d'occurrence. Les résultats expérimentaux réalisés pour plusieurs textes ont démontré l'intérêt de leur méthode d'indexation.

[MoHA12] ont proposé et implémenté une méthode pour indexer des livres arabe en utilisant l'analyse syntaxique. Le processus dépend en grande partie du processus de synthèse et d'abstraction du texte pour collecter automatiquement les principaux sujets et déclarations du livre.

Cette approche donne de bons résultats dans des situations spécifiques, telles que la détermination du sens exact d'un mot ambigu tel qu'il est exprimé dans la phrase, mais reste moins efficace que d'autres approches, compte tenu de la complexité de la langue arabe.

5.2.1.2. L'approche statistique

L'approche statistique repose principalement sur des techniques statistiques. Une variété de cette approche a été développée pour extraire des descripteurs et étudier leur apparition dans un document, voire dans le corpus.

La distribution de fréquence des mots a été un objet d'étude clé dans l'approche statistique au cours des dernières décennies. Cette distribution suit approximativement une forme mathématique simple appelée « loi de Zipf ». Selon cette loi, les mots apparaissent selon une distribution de fréquence systématique, de sorte qu'il existe peu de mots à très haute fréquence représentant la majeure partie du texte et de nombreux mots à basse fréquence. Nous mentionnons très brièvement quelques-uns des endroits où cette loi affecte la recherche dans notre étude :

- La « loi de Zipf » montre combien de texte les doit examiner et à quel point les statistiques doivent être précises pour atteindre le niveau d'erreur attendu [Finc93].
- La « loi de Zipf » fournit également un modèle de base pour l'occurrence attendue de termes cibles et les réponses à certaines questions peuvent fournir des informations considérables sur son rôle dans le corpus [StPo98] : Que signifie demander si un mot est significatif dans un corpus, au-delà de la simple occurrence ou de la probabilité relative ? Quelle est l'étendue de l'influence sémantique d'un mot dans un corpus ? En quoi le modèle d'occurrences contribue-t-il à notre évaluation de sa pertinence dans le corpus ?[Powe98].
- La loi de Zipf fournit une base pour évaluer les analyseurs syntaxiques et les tagueurs [EnPo98]. Encore une fois, nous résumons le rôle potentiel sous la forme d'une série de questions : Comment un modèle de langage développé sur un corpus est-il transféré à un autre ? Comment traduire les estimations de performance de quelques corpus testés en estimations pour la langue dans son ensemble ? Comment les différences de registre, de genre et de support affectent-elles l'utilité d'un système et comment compensons-nous ces différences ? [Powe98].

Les approches statistiques reposent principalement sur des techniques statistiques. Une variété de ces approches a été développée pour extraire des descripteurs (termes) et étudier leur apparition dans un document, voire dans le corpus. La méthode « Term Frequency–Inverse Document Frequency : $tf - idf$ » est l'une des méthodes statistiques qui fournit une bonne représentation du poids des mots de corpus dont la taille du document est homogène. Plusieurs alternatives ont été proposées pour la méthode $tf - id$, qui a fait l'objet de nombreuses études comparatives.

La faisabilité de cette approche dépend également du processus d'extraction de la racine/tige de chaque mot, selon une approche basée sur la racine ou sur le lemme, afin de surmonter l'ambiguïté du mot (polymorphisme du mot).

Plusieurs études ont montré que le processus consistant à extraire le mot de ses préfixes et suffixes est plus utile, pour les systèmes de recherche d'informations en arabe, que d'autres approches.

Les chercheurs ont adopté diverses méthodes et techniques statistiques dans le processus d'indexation [ElIs06] [Khre06] [Elha15] [Thab08] [AlAA08] [TEZH09] [GhHF09] [AlKG06] [RaDi10].

En conclusion, ces méthodes, considérées comme simples à mettre en œuvre, sont efficaces et tolèrent parfaitement les grandes masses documentaires. D'autre part, l'hypothèse considérant les mots comme des unités indépendantes génère une perte d'informations sémantiques. Les index résultants peuvent générer des problèmes d'ambiguïté et s'écarter du contexte général du document [BSZM16].

5.2.1.3. L'approche sémantique

Cette approche vise, d'une part, à réduire l'ambiguïté du sens des mots et, d'autre part, à extraire les relations sémantiques entre ces mots. Ainsi, les textes se concentrent sur l'unité de signification plutôt que de simples mots. Les relations sémantiques peuvent également être calculées à l'aide de méthodes permettant d'évaluer la quantité d'informations entre les mots.

[THYB07] ont intégré le processus sémantique dans un moteur de recherche Internet et ont utilisé plusieurs techniques (Harman, Croft et Okapi) pour évaluer les performances de ce moteur. Dans une étude récente, [ADAC16b] [AEAC13] ont exploité la base lexicale de WordNet arabe dans un IRS afin d'indexer la collection de documents et la requête de l'utilisateur. D'autres [MRRZ14] ont introduit une approche d'expansion des requêtes utilisant

une ontologie construite à partir de pages Wikipédia en plus d'autres thésaurus pour améliorer la précision de la recherche en arabe.

Cette approche offre la meilleure couverture sémantique pour les documents car elle repose sur des ressources sémantiques (dictionnaires, anthologies ou autres). Cependant, elle reste limitée par le type de ressource utilisé et sa capacité à décrire les mots du texte en cours de traitement.

5.2.1.4. L'approche hybride

Plusieurs chercheurs [HaEA11] [MoWa10] [AlAb17] [DiBe12] ont expérimenté différentes combinaisons de méthodes linguistiques, statistiques et sémantiques, tirant parti des avantages de chaque méthode pour tenter de surmonter leurs lacunes et d'améliorer le processus d'indexation en extrayant informations cachées dans un document. Ces approches ont souvent conduit à de meilleurs résultats que ceux obtenus par l'utilisation de méthodes standard.

Bien que les résultats de cette approche soient positifs, elle souffre du problème de la complexité, en fonction de l'intégration d'autres approches.

5.2.2. L'extraction de mots-clés arabes

Les mots-clés (descripteurs) sont un sous-ensemble de mots ou d'expressions pouvant décrire la signification d'un document ; plusieurs applications de traitement du langage naturel pouvant en tirer parti. Malheureusement, la plupart des auteurs ne désignent pas ces mots dans une partie spécifique de ses documents. D'autre part, l'ajout manuel de mots clés de haute qualité est coûteux, prend du temps et génère des erreurs. Par conséquent, ce domaine a émergé pour développer de nouveaux algorithmes et systèmes conçus pour extraire les mots-clés automatiquement.

[ElRa09] ont présenté le système KP-Miner (Keyphrases-Miner) permettant d'extraire des mots clés (phrases) à partir de documents anglais et arabes de longueur variable. Ce système n'a pas besoin d'être formé à un jeu de documents particulier pour pouvoir remplir sa tâche (apprentissage non supervisé). Il présente également l'avantage d'être configurable, car les règles et les méthodes heuristiques adoptées par le système sont liées à la nature générale des documents et des mots clés. En général, des expériences et des études comparatives avec des systèmes largement utilisés suggèrent que KP-Miner est efficace.

[AmFo16] ont introduit AKEA (an Arabic Keyphrase Extraction Algorithm), un algorithme d'extraction de mots clés - non supervisé - pour les documents arabes simples. Ils

se sont appuyés sur des méthodes heuristiques permettant de créer des modèles linguistiques fondés sur des balises POS (Part-of-Speech), des connaissances statistiques et le modèle structurel interne des termes. Ils ont utilisé Wikipedia arabe pour améliorer le classement des mots clés candidats en ajoutant un score de confiance si le candidat existe sous la forme d'un concept Wikipédia indexé. Les résultats expérimentaux ont montré que les performances d'AKEA sont supérieures à celles des autres algorithmes non supervisés, car ils ont fourni des valeurs de précision plus élevées.

[AAAW13] ont présenté un système d'extraction de mots clés pour les documents arabes en utilisant des informations statistiques de terme de cooccurrence. Dans le cas où la cooccurrence d'un terme est dans le degré de biais, alors le terme est important et il est probable qu'il s'agisse d'un mot clé. Le degré de biais des termes et l'ensemble des termes fréquents sont mesurés à l'aide de test du (χ^2)[32]. Par conséquent, les termes avec des valeurs χ^2 élevées sont susceptibles d'être des mots-clés. Cette technique a montré une performance acceptable par rapport à d'autres techniques.

[ElAl12] ont présenté une technique d'apprentissage supervisé permettant d'extraire des mots clés de documents arabes. L'extracteur est doté de connaissances linguistiques lui permettant d'accroître son efficacité au lieu de se fier uniquement à des informations statistiques telles que la fréquence des termes et la distance. Un corpus arabe annoté est utilisé pour extraire les caractéristiques lexicales requises des mots du document. La connaissance comprend également des règles syntaxiques basées sur des balises POS et permet aux séquences de mots d'extraire les mots clés candidats. Les expériences réalisées montrent l'efficacité de cette méthode pour extraire des mots clés arabes.

[DuHe16] ont présenté un cadre permettant d'extraire des mots clés de documents d'actualité en arabe. Il s'appuie sur l'apprentissage supervisé, les Naïfs Bayes en particulier, pour extraire les mots clés. Le dernier ensemble de mots clés est choisi parmi l'ensemble des mots ayant une probabilité élevée d'être des mots clés.

Diverses expériences ont montré l'efficacité de ces méthodes pour extraire des mots-clés arabes avec des pourcentages variables. Cependant, alors que les techniques supervisées sont

[32] Un test du khi-carré, également appelé test 2, est un test d'hypothèse statistique où la distribution d'échantillonnage de la statistique de test est une distribution du khi-carré lorsque l'hypothèse nulle est vraie. Sans autre qualification, le « test du chi-carré » est souvent utilisé comme test abrégé pour le test du chi-carré de Pearson. Le test du khi carré permet de déterminer s'il existe une différence significative entre les fréquences attendues et les fréquences observées dans une ou plusieurs catégories.

coûteuses et limitées par le type de ressources linguistiques utilisées, les techniques non supervisées souffrent d'une faible couverture sémantique des documents.

5.3. Caractéristiques de la langue arabe

Les caractéristiques grammaticales et morphologiques complexes de la langue arabe rendent la tâche du traitement automatique plus difficile, et plus précisément, le traitement sémantique. Parmi ces caractéristiques, en particulier celles liées indirectement au traitement sémantique, nous soulignons les suivantes :

- Les écritures arabes ont des signes diacritiques pour représenter les voyelles courtes, qui sont des marques au-dessus ou au-dessous des lettres. Cependant, ces diacritiques ont disparu de la plupart des écrits contemporains et les lecteurs sont censés combler les diacritiques manquants grâce à leur connaissance de la langue. L'absence de signes diacritiques dans les textes arabes contemporains rend le traitement automatique difficile.
- L'analyse morphologique est une procédure complexe car l'arabe est une langue agglutinante. Par exemple, le mot "أفاستسقيناكموها" (est-ce que nous vous avons demandé - de l'eau au pluriel - de l'eau pour elle) est l'un des mots les plus longs de la langue arabe. Il se compose de 15 lettres et 9 diacritiques. Sa racine est le verbe "سقى" (arroser). Nous ajoutons au mot le préfixe "است" pour devenir "استسقى" (il a demandé de l'eau). En ajoutant un pronom sujet, le mot devient "استسقينا" (nous avons demandé de l'eau). Ensuite nous ajoutons le pronom d'objet indirect pour devenir "استسقيناكم" (nous vous avons demandé - pluriel- pour de l'eau), et nous avons ajouté l'objet direct pour devenir "استسقيناكموها" (nous vous avons demandé - de l'eau au pluriel - de l'eau pour elle), puis nous ajoutons "F" d'appel (ف الاستئناف) et "A" de question (أ الاستفهام) pour devenir une expression tout à fait significative : "أفاستسقيناكموها" (est-ce que nous vous avons demandé - de l'eau au pluriel - de l'eau pour elle).
- L'arabe est une langue très flexionnelle et dérivationnelle où de nombreux noms et verbes sont dérivés de la même racine. Ce dernier est basé sur plus de 150 patterns (ou patrons), ce qui les rend plus complexes et difficiles à traiter.

5.4. Le système d'indexation semi-automatique

Comme nous avons souligné dans l'introduction, nous avons conçu et développé un système d'indexation semi-automatique basé sur :

- Une indexation semi-automatique en ligne de documents textuels arabes (Figure 5.1).
- Une indexation automatique hors ligne du corpus arabe (Figure 5.5).

5.4.1. Système d'indexation semi-automatique en ligne

Ce système comprend trois unités : une unité d'indexation automatique, une unité d'extraction automatique des mots-clés et une unité de mise à jour d'un index partiel d'un document après l'intervention de l'expert humain pour la sélection des mots-clés pertinents.

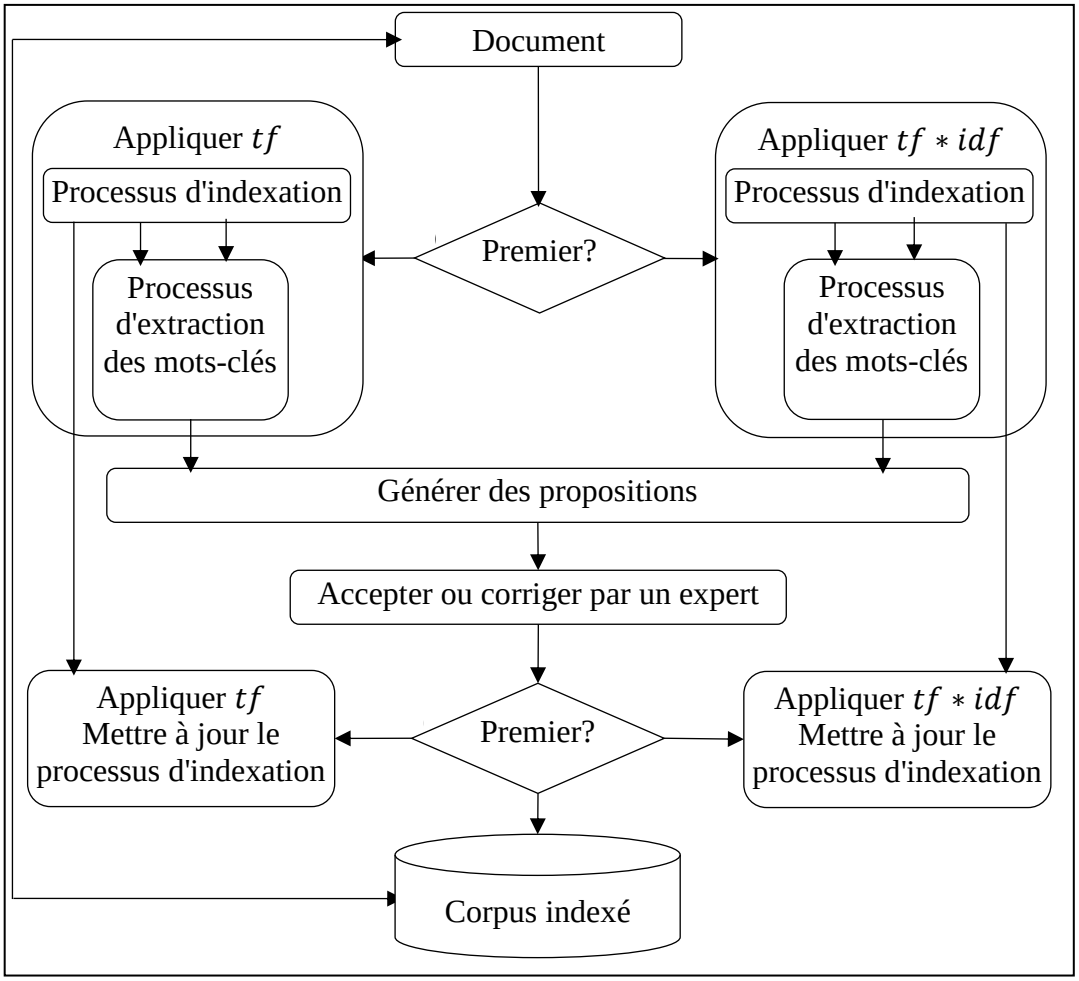

Figure 5.1: Système d'indexation semi-automatique en ligne de documents textuels arabes

De plus, nous avons intégré notre système d'indexation en ligne à un éditeur de texte arabe (Figure 5.6) que nous avons conçu et implémenté dans le but d'effectuer nos expériences. Nous avons également créé un corpus arabe dans un nouveau format (Figure 5.7) qui nous permet d'effectuer les expériences nécessaires.

5.4.1.1. Unité d'indexation automatique

L'indexation est le processus de représenter un texte donné dans la liste des termes informatifs, en vue d'en faciliter le repérage et la consultation.

L'indexation automatique des textes arabes a dominé la plupart des travaux de recherche d'information de texte arabe. Dans notre étude, et comme nous avons vu dans le chapitre précédent, nous avons suivi l'approche [DiBe12] pour créer l'index avec quelques modifications. Cette méthode s'est révélée efficace pour améliorer le processus d'indexation des documents arabes.

5.4.1.1.1. Encodage

Le corpus et les requêtes peuvent être codés différemment, ce qui les rend incomparables. Afin de normaliser les documents avec les requêtes, nous devons convertir le tout en codage UTF-16, car il permet la représentation de lettres et de symboles dans un large éventail de langues, y compris l'arabe.

5.4.1.1.2. Normalisation, les mots vides et la lemmatisation

Nous appliquons les mêmes étapes de normalisation, suppression des mots vides et lemmatisation vues dans le chapitre précédent.

5.4.1.1.3. Fréquence du terme et pondération

Dans notre étude, nous avons utilisé le $tf - idf$ qui combine les définitions de fréquence du terme et de fréquence inverse de documents pour produire un poids composite pour chaque terme de chaque document. La procédure de pondération $tf - idf$ attribue un poids au terme t dans le document d donné par :

$$tf - idf_{t,d} = tf_{i,j} * idf_i \qquad \text{(Eq.5.1)}$$

- $tf_{i,j}$: le nombre de fois que ce terme i apparaît dans le document j.
- $idf_i = \log \frac{|\mathrm{D}|}{|\{d_i : t_j \in d_j\}|}$
- $|\mathrm{D}|$: nombre total de documents dans le corpus.

- $|\{d_i : t_j \in d_j\}|$: nombre de documents où le terme t apparaît (c.-à-d., $tf(t,d) \neq 0$.

Notre unité d'indexation automatique traite différemment le premier document ajouté au corpus (Figure 5.2). Comme il n'y avait pas de documents disponibles avant le premier document pour calculer $tf - idf_{t,d}$, nous ne comptons que la valeur $tf_{i,j}$.

L'unité d'indexation automatique construit un index partiel pour chaque document de chaque corpus. Si non, la sortie de cette unité est un *index partiel* pour chaque document (Figure 5.2). La principale motivation derrière la construction d'index partiels est de permettre à l'expert d'intervenir plus tard dans la création d'index.

Indexing function pseudo code

```
Input: Document di ∈ corpus
Output: Indexi // partial index
begin
      For each token in di loop
            Encoding ();
            Normalize ();
            Removing_stop_words ();
            Stemming ();
            If (tf_type = tf) then
                  Weighting(tf)
            Else
                  Weighting(tf-idf);
            End
            Stored tf for the term = token
      End loop.
      Add di to Indexi.
end.
```

Figure 5.2: Algorithme d'indexation automatique

5.4.1.2. Unité d'extraction automatique de mots clés

Nous avons adopté une méthode simple d'extraction de mots-clés, dans la mesure où l'expert humain est responsable de la décision finale concernant l'acceptation ou la modification des mots-clés appropriés pour le document en cours de traitement (voir l'exemple de la Figure 5.3).

Instructions	Execute?	If no, why?
1. ***Input:*** في تقريرها السنوي حول الجمهورية الجزائرية الديمقراطية الشعبية أن ... (In its annual report on the democratic and popular republic of Algeria that …)	-	
2. ***Selected word from the result of the indexing module*** في تقريرها السنوي حول الجمهورية [الجزائرية] الديمقراطية الشعبية أن...	Yes	
3. ***Add 1st right word*** في تقريرها السنوي حول [الجمهورية الجزائرية] الديمقراطية الشعبية أن...	Yes	
4. ***Add 2nd right word*** في تقريرها السنوي [~~حول~~ الجمهورية الجزائرية] الديمقراطية الشعبية أن...	No	Stop word
5. ***Add 1st left word*** في تقريرها السنوي حول [الجمهورية الجزائرية الديمقراطية] الشعبية أن...	Yes	
6. ***Add 2nd left word*** في تقريرها السنوي حول [الجمهورية الجزائرية الديمقراطية الشعبية] أن...	Yes	
7. ***Output:*** الجمهورية الجزائرية الديمقراطية الشعبية (The democratic and popular republic of Algeria)		

Figure 5.3: Exemple d'extraction automatique de mots clés

L'unité d'extraction automatique de mots-clés (Figure 5.4) propose la liste des mots candidats. Cette liste est limitée à douze mots-clés, chacun composé d'au plus cinq mots. Ces mots sont extraits en deux étapes :

Dans la première étape, nous adoptons les résultats de l'unité d'indexation automatique, dans laquelle nous récupérons les mots d'index avec les poids les plus élevés. Ensuite, nous ajoutons, si possible, à chaque mot d'index, à partir du texte original, deux mots les plus proches voisins à droite et deux autres à gauche, tout en veillant à ce que cette chaîne de cinq mots ne contienne pas de signes de ponctuation arabes. Sinon, nous prenons simplement le nombre de mots entre deux ponctuations. Nous donnons également la priorité à une phrase nominale en définissant les termes pour les mots candidats dans l'ordre suivant :

- Les mots commençant par les lettres « ال » et finissant par « ي », « ة » ou « ء ».
- Les mots commençant par les lettres « ال ».
- Les mots qui se terminent par les lettres « ي », « ة » ou « ء ».
- Les mots ordinaires.

```
     Keywords_Extract function pseudo code

     Input: Document di ∈ corpus

Output: Keywords [ ]
Begin
    For j = 1 to 12 loop
        word ← Paratial_Index.canditat_word[j];
        word ← From_Original_text (word);
        if (Setting_terms (fst_right_word))
            word ←   word + fst_right_word;
        if (Setting_terms (snd_right_word))
            word ←   word + snd_right_word;
        if (Setting_terms (fst_leftt_word))
            word ←   fst_left_word + word;
        if (Setting_terms (snd_lest_word))
            word ←   snd_left_word + word;
       Keywords [i] ← word
    End loop.
End.
```

Figure 5.4: Algorithme d'extraction automatique de mots clés

Dans la deuxième étape, nous proposons à l'expert humain douze mots clés classés par ordre décroissant, ensuit l'expert humain accepterait ou modifierait les suggestions générées par l'unité d'extraction automatique de mots clés.

5.4.1.3. Unité de mise à jour d'index partiel

Le rôle de cette unité est de mettre à jour un *index partiel* d'un document. Les opinions de l'expert sont prises en compte en mettant à jour les poids des mots d'index sélectionnés et en leur attribuant des valeurs plus élevées. Cette phase se termine par l'intégration de cet *index partiel* dans le document et par sa sauvegarde dans un fichier objet afin de l'exploiter ultérieurement.

5.4.2. Système d'indexation hors ligne pour la génération et la mise à jour d'index général

Le rôle de ce système est de générer et de mettre à jour un *index général* basé sur des *index partiels* de plusieurs corpus (Figure 5.5).

Il récupère tous les index de documents (index partiels) créés par le système d'indexation semi-automatique en ligne et les fusionne dans un seul index général. Il met également à jour cet index chaque fois que nécessaire.

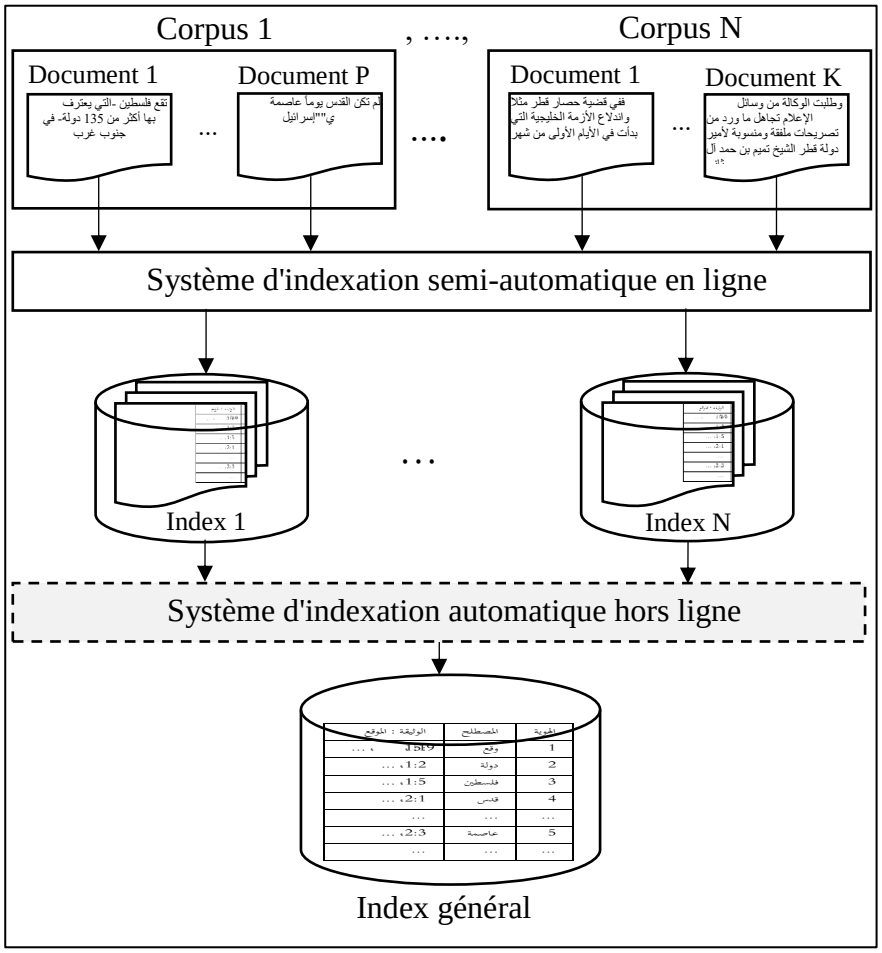

Figure 5.5: Système d'indexation automatique hors ligne

5.5. Applications implémentées

Pour mettre en œuvre le système d'indexation semi-automatique en ligne que nous avons conçu, nous avons développé un éditeur de texte arabe contenant un système d'indexation de documents en ligne. En outre, nous avons travaillé à la création d'une nouvelle forme de corpus arabe, contenant les mots-clés proposés par un expert humain, afin de mener les expériences nécessaires. Nous avons également utilisé l'application OIRDA pour l'indexation générale et la recherche d'informations et nous l'avons doté d'un système d'indexation automatique hors ligne permettant de générer et de mettre à jour l'*index général*.

5.5.1. Éditeur de texte arabe

Nous avons d'abord développé un éditeur de texte arabe (Figure 5.6), qui, en plus des fonctions habituelles d'éditeur de texte, est fourni avec l'option d'indexation automatique des utilisateurs de l'éditeur. Nous avons adopté la conception du système d'indexation semi-automatique en ligne décrite ci-dessus (Figure 5.1) pour ajouter cette option.

Figure 5.6: L'éditeur de texte arabe « SIRAT »

Comme discuté ci-dessus, nous traitons différemment le premier document ajouté au corpus, où il n'y a pas d'autres documents, de sorte qu'il ne compte que la valeur $tf_{i,j}$. Nous intégrons ensuite l'unité d'extraction de mots-clés, qui repose sur les résultats obtenus de l'unité d'indexation automatique en proposant des suggestions de mots-clés aux indexeurs experts, leur permettant ainsi de modifier les mots proposés. Enfin, l'index est mis à jour. La sortie de cet éditeur est un fichier objet contenant le texte traité et l'*index partiel* généré.

5.5.2. Nouvelle forme de corpus arabe

Pour étudier l'efficacité du système proposé, il était nécessaire d'obtenir un corpus de test constitué d'un ensemble de documents arabes répondant à un ensemble de fonctionnalités nécessaires et suffisantes pour le test.

Nous avons développé un programme pour construire un corpus arabe, en organisant un certain nombre de pages Web du site Web d'Al-Jazeera[33], sous une nouvelle forme de corpus

[33] http://www.aljazeera.net/encyclopedia. Téléchargé le 16 November 2017.

différente de celle habituelle, en ajoutant des mots-clés suggérés par l'expert humain (les journalistes d'Al-Jazeera) à la fin des documents (Figure 5.7). Cela permet d'évaluer les performances de l'unité d'extraction automatique de mots-clés. De plus, nous avons pris en compte l'ensemble des règles utilisées globalement dans la construction de ce corpus, en particulier celles fournies par (TREC) [LaCo01b].

<DOC>

<DOCNO> ALJAZEERA-511 </DOCNO>

<HEADLINE> ماذا تعرف عن الجيوش الإلكترونية؟ </HEADLINE>

<DATELINE> المصدر: الجزيرة + وكالات, مواقع إلكترونية </DATELINE>

<TEXT>

الجيوش الإلكترونية مجموعات مدربة تعمل وفق أجندة خاصة هدفها اختراق مواقع الخصوم، والترويج لوجهة نظر معينة عبر مختلف منصات الإنترنت، وإسكات وتشويه سمعة المناوئين، إلى جانب ترويج الإشاعات والأكاذيب وخلق البلبلة، وقد بدأت الدول في إنشاء وحدات إلكترونية داخل أجهزتها العسكرية والأمنية لحماية أمنها القومي. خدمة رسمية الجيوش الإلكترونية مجموعة من الأشخاص وقراصنة الإنترنت (هاكرز) تعمل لصالح أجهزة المخابرات والأمن في الغالب، تسعى لاختراق المواقع الإلكترونية الخاصة بالشخصيات والمؤسسات والدول، ولا تكاد تترك منتديات أو نقاشات أو تعليقات على مواقع التواصل الاجتماعي وغيرها من المواقع الإلكترونية إلا ودخلت إليها للدفاع عن وجهة النظر الرسمية، ونشر الإشاعات والأكاذيب التي تربك رؤية الناس وتوجههم باتجاه معين.

…

وفي عصر المعلومة والحروب الإلكترونية، بدأت دول عديدة سياسة إنشاء "جيوش إلكترونية" نظامية لها ميزانيتها الخاصة، وتسعى للدفاع عن البلاد ضد الهجمات الإلكترونية التي لا تكاد تنتهي حتى تبدأ. وقد أعلنت وزيرة الدفاع الألمانية أورسولا فون دير لاين يوم 6 أبريل/نيسان 2017 عن تكوين جيش إلكتروني كوحدة مستقلة داخل الجيش الألماني إلى جانب القوات البرية والبحرية والجوية، حيث يمارس مهام دفاعية وهجومية على شبكة الإنترنت. وقالت الوزيرة إن عمل الجيش الإلكتروني لن يقتصر على صد هجمات القرصنة، بل سيرد عليها أيضا في ساحة المعركة، وهي الإنترنت. وأضافت "في حال تعرض شبكات الجيش الألماني للهجوم فمن حقنا أيضا أن نرد".

</TEXT>

<KEYWORDS> الجيش الإلكتروني؛ إنترنت؛ حروب إلكترونية؛ فيروسات؛ الجيش الإلكتروني السوري؛ موسوعة الجزيرة؛ </KEYWORDS>

</DOC>

Figure 5.7: Un exemple sur la nouvelle forme de corpus arabe

Ainsi, nous avons pu obtenir un corpus arabe contenant 2416 documents et 25 requêtes. Le nombre de vocabulaire de ce corpus est de 1475148 mots, dont 133474 mots différents (soit 9.03% du total des mots).

Selon la loi de Zipf, qui concerne la distribution des mots dans le document, la portée et l'importance des mots de corpus. La Figure 5.8 illustre la courbe du corpus d'Al Jazeera (en rouge représentée par les symboles (+)) et la courbe de Zipf (en vert représentée par les symboles (x)). La figure montre que la courbe du corpus d'Al Jazeera est très proche de la courbe de Zipf. En outre, selon certains autres critères [KaNi06b], notre nouvelle forme de corpus est très riche et qualifiée pour être utilisée comme corpus d'évaluation des systèmes de recherche d'informations.

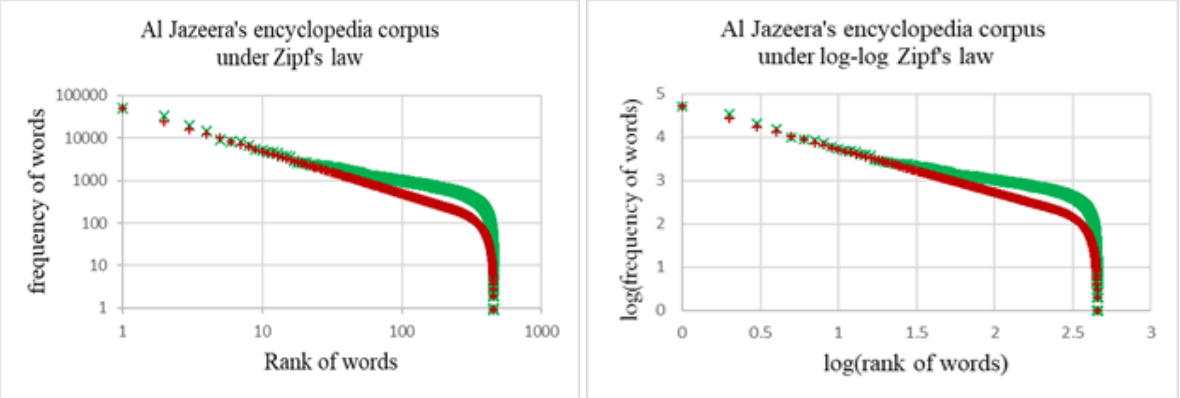

Figure 5.8: Courbe de corpus de site Al Jazeera selon la courbe de loi de Zipf

Cette nouvelle forme nous permet de bénéficier, entre autres :

- La contribution à la construction d'un système d'évaluation des IRS, qui permet aux chercheurs de tester l'efficacité de leurs applications. Outre la qualité et la quantité des documents pris en compte dans ce corpus, nous avons créé deux types de requêtes et leurs documents correspondants. Le premier est un bref et simple ; tandis que le second est vaste et complexe, basé sur les mots-clés du corpus, par exemple : "الحرب الالكترونية التي يقودها الجيش السوري" (Guerre électronique dirigée par l'armée syrienne).
- La contribution à la construction d'un système d'évaluation des systèmes d'extraction de mots-clés, où nous avons pu effectuer des expériences d'extraction à l'aide des documents du corpus, comparer les résultats de ces systèmes avec les mots-clés disponibles et calculer les scores de précision et de rappel.

5.6. Analyse et résultats

Nous avons mené une série d'expériences pour montrer l'effet de chaque méthode d'indexation sur les performances d'extraction, et d'évaluer la performance de différentes méthodes d'indexation dans la recherche d'informations en arabe, en utilisant l'application OIRDA dotée d'un système d'indexation hors ligne pour l'indexation générale et la recherche.

Nous comparons d'abord les deux modèles d'indexation suivants :

- Indexation basée sur les mots clés (*Keyword-based indexing*) : l'index est composé uniquement de mots clés approuvés par l'expert.
- Indexation sans mots clés (without keyword-based) ou normale indexation : l'index est généré par une unité d'indexation automatique sans l'intervention de l'expert.

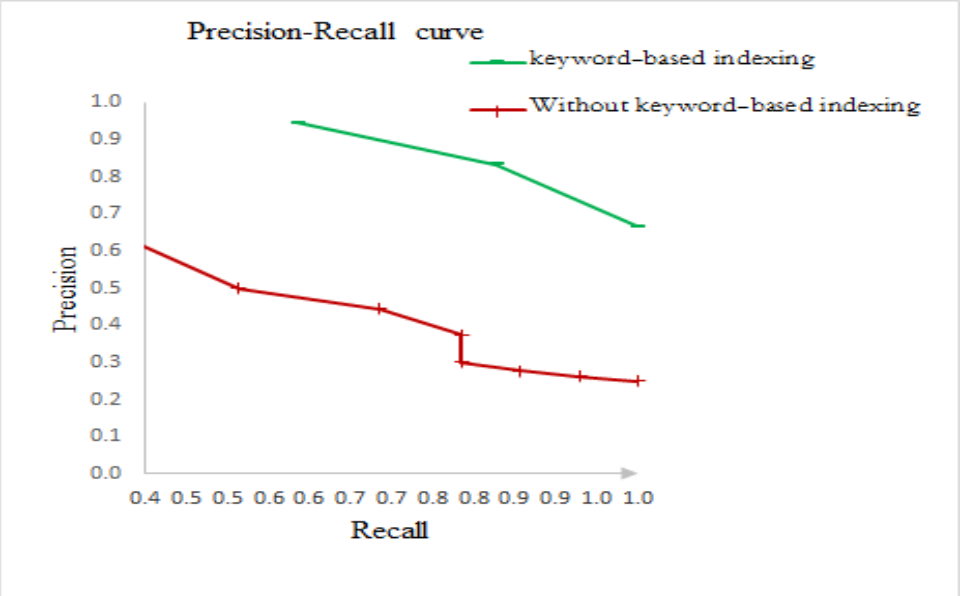

Figure 5.9: Comparaison entre l'indexation basée sur les mots clés et sans mots clés

La Figure 5.9 représente une comparaison entre ces deux modèles en fonction de leurs courbes de précision et rappel. Les résultats montrent que le modèle d'indexation basé sur les mots-clés, la courbe en rouge représentée par des symboles (-), est plus efficace que le modèle d'indexation sans mots-clés, la courbe en vert représenté par des symboles (+), sur tous les points de rappel et de précision.

Ensuite, nous comparons les deux modèles d'indexation suivants :

- Hybride : différentes combinaisons d'indexation basées sur mot-clé et d'indexation sans mot-clé, de manière à exprimer les avantages de chacune d'elles.
- Indexation basée sur les mots clés (*Keyword-based indexing*).

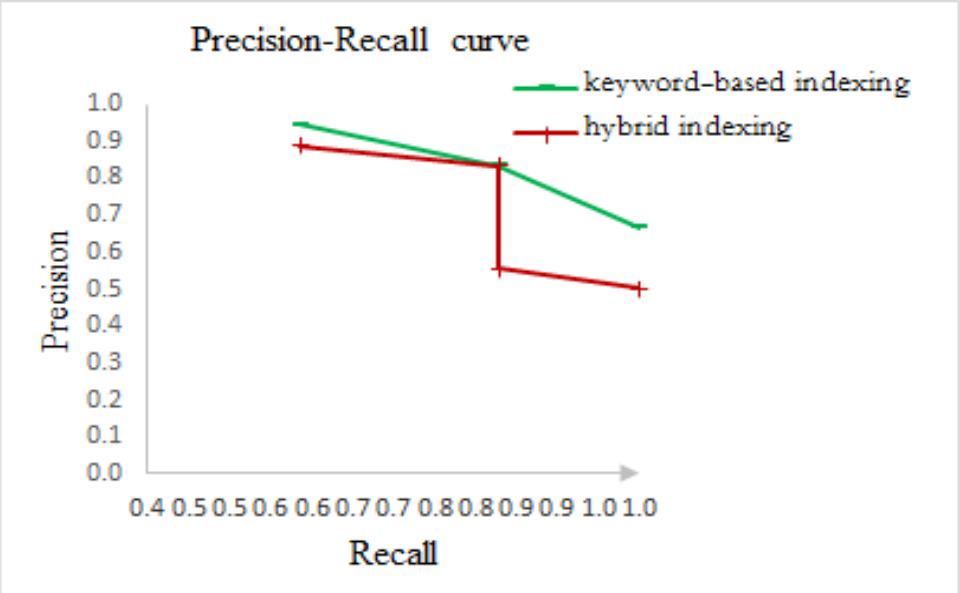

Figure 5.10: Comparaison entre l'indexation hybride et basée sur les mots clés

Dans la série de nos expériences, les résultats montrent que le modèle basé sur les mots clés, la courbe de couleur verte représentée par les symboles (-), est plus efficace que le modèle hybride, la courbe de couleur rouge représentée par les symboles (+). On peut observer ce comportement dans la Figure 5.10 ; l'indexation basée sur des mots-clés représentée par la courbe précision / rappel est supérieure à celle de l'indexation hybride.

En outre, les résultats montrent que le modèle d'indexation basé sur des mots clés est la meilleure approche car elle est plus efficace pour identifier les descripteurs les plus pertinents pour le document. Cela est principalement dû à l'intervention de l'expert humain dans l'identification des mots clés, en particulier lors de requêtes ambiguës comprenant une polysémie, des mots composés, etc., qui nécessitent un traitement sémantique précis.

Ce modèle s'est également révélé efficace pour réduire au minimum la taille de la mémoire de stockage et ainsi améliorer le temps de réponse des différentes requêtes.

Cependant, le modèle d'indexation basé sur les mots clés souffre d'un inconvénient majeur : dans le cas où l'expert ne peut pas identifier les descripteurs les plus pertinents pour le document, l'aspect que ce modèle doit améliorer et trouver une solution viable.

5.7. Conclusion

L'objectif principal de cette étude est de montrer les effets de l'indexation en ligne, qui nécessite l'indexation semi-automatique, sur les performances du système d'extraction d'informations. En outre, ce modèle s'est avéré efficace pour minimiser la taille de la mémoire de stockage des index et ainsi améliorer le temps de réponse des différentes requêtes. Par conséquent, nous recommandons d'intégrer ce modèle dans des outils de traitement de texte afin de permettre à l'éditeur de contribuer efficacement à la création d'index de haute qualité tout en prenant en compte les inconvénients de ce modèle. Cette étude propose également une solution aux problèmes et carences dont souffre le traitement de la langue arabe, notamment en ce qui concerne la création de corpus, en développant un cadre d'application pour la création et le développement de corpus. En outre, l'étude suggère une solution pour réduire les carences des systèmes d'évaluation de systèmes de la recherche d'informations, qui permet aux chercheurs de tester leurs algorithmes d'indexation et de recherche.

Dans ce chapitre suivant, nous allons présenter la démarche suivie pour préparer la nouvelle structure de ressource lexicale et la problématique de la mesure de similarité sémantique afin de montrer l'efficacité de cette nouvelle ressource lexicale à la désambiguïsation. Nous allons présenter également l'algorithme global utilisé au cours du processus de la désambiguïsation sémantique des mots arabes.

Chapitre 6 :

Mesure de similarité sémantique locale et algorithme global pour la désambiguïsation des sens des mots arabes, basé dictionnaire contemporaine

6. MESURE DE SIMILARITE SEMANTIQUE LOCALE ET ALGORITHME GLOBAL POUR LA DESAMBIGUÏSATION DES SENS DES MOTS ARABES, BASE DICTIONNAIRE CONTEMPORAINE

6.1. Introduction

L'ambiguïté (ou le problème de la polysémie [KoMa00]) est la propriété d'un mot, voire d'une suite de mots, ayant plusieurs significations ou plusieurs analyses grammaticales possibles, et ayant un effet négatif sur les performances des systèmes de recherche d'informations textuelles (RI). Cependant, la plupart des systèmes de RI traditionnels écartent totalement le traitement de la « polysémie », bien que les documents contiennent uniquement les sens pertinents des mots d'une requête particulière. Cela améliorerait sans aucun doute la précision des systèmes de RI.

La désambiguïsation sémantique des mots (Word Sense Disambiguation : WSD) vise à déterminer le sens approprié pour un mot ambigu apparu dans un contexte précis [AgEd07] [Navi09], en permettant une meilleure compréhension, et par la suite, une meilleure représentation des contenus textuels. C'est une tâche essentielle pour les applications de la recherche d'information sémantique.

Au cours ces dernières années, la performance des systèmes de désambiguïsation sémantique des mots (WSD) a augmenté. Cependant, Il existe peu de travaux de WSD sur le texte arabe par rapport à d'autres langues, malgré les efforts considérables déployés pour répondre aux besoins des applications de traitement automatique du langage naturel.

La langue arabe présente plusieurs défis pour WSD, cela est dû d'une part à ses caractéristiques linguistiques, tels que le manque de diacritiques pour les textes arabes contemporains, parce qu'un mot sans diacritiques augmente le nombre des sens possibles par rapport au même mot avec diacritiques, et la richesse morphologique de la langue arabe augmente le taux de l'ambiguïté dans les textes. D'autre part, le manque de ressources lexicales nécessaires présente un majeur défi pour WSD basée sur l'utilisation de la ressource de connaissance.

Dans ce chapitre, nous présentons un algorithme de WSD basé sur sens (Sene-based WSD) en utilisant une sémantique initiale et une autre supplémentaire sous la forme des mots liés sémantiquement dérivés des définitions et des exemples de la ressource de lexicale

« dictionnaire de la langue arabe contemporaine ». Cet algorithme prend en compte les informations liées au domaine associé aux mots du document.

Dans une autre approche de WSD décrite dans ce chapitre, nous utilisons le sens de voisins d'un mot. Cette approche basée sur le contexte (Context-based WSD) inclut les informations contextuelles dans WSD pour trouver le sens approprié pour les mots indexés des documents et des requêtes. Les mots dans la phrase (mots voisins) sont appelés mots de fenêtre de contexte. Cet algorithme utilise les mots sémantiquement communs des mots de la fenêtre contextuelle.

Dans ce chapitre, nous commençons par présenter la démarche suivie pour préparer la nouvelle structure de ressource lexicale « DiLAC[34] » pour être exploitable. Ensuite, nous abordons la problématique de la mesure de similarité sémantique afin de montrer l'efficacité de notre ressource lexicale à la désambiguïsation, en analysant des résultats des expériences réalisées. Nous présentons également l'algorithme global utilisé au cours du processus de la désambiguïsation sémantique des mots arabes.

6.2. Le modèle proposé

Nous proposons un modèle de recherche d'information sémantique qui se compose essentiellement d'un module WSD (Figure 6.1), ce module consiste en deux composants : (1) la ressource lexicale « DiLAC » où nous avons vu ses caractéristiques puissantes dans le chapitre 3, et sur laquelle nous basons pour extraire le sens de chaque lemme à désambiguïser, après que nous la préparons et l'évaluons qu'elle sera apte à exploiter, en utilisant l'algorithme de similarité sémantique locale ; (2) et le processus de WSD, qui utilise un algorithme global pour la désambiguïsation sémantique des mots arabes, basant sur le dictionnaire cité.

[34] DILAC : **Di**ctionnaire de la **L**angue **A**rabe **C**ontemporaine.

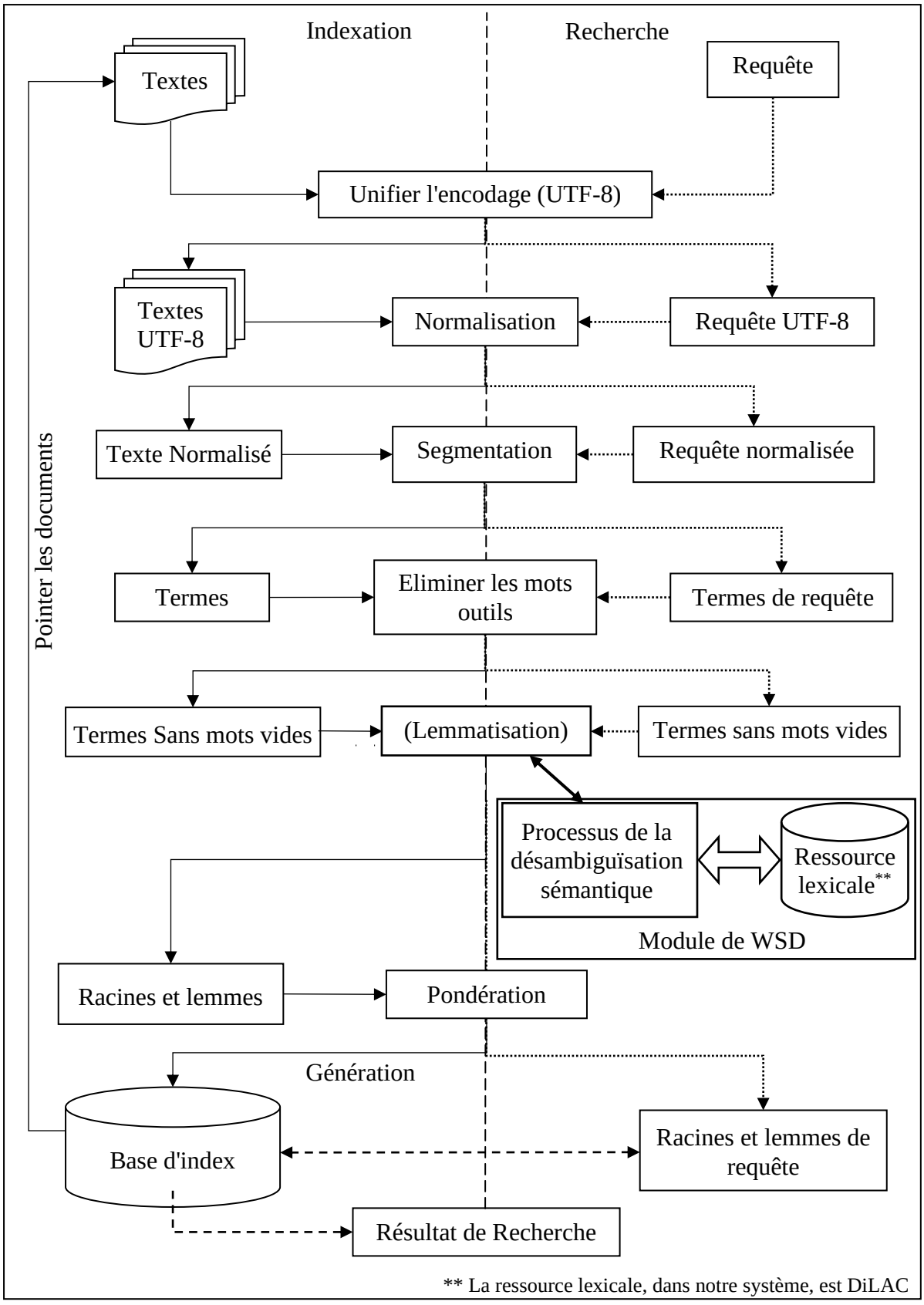

Figure 6.1: Modèle proposé pour la recherche d'information sémantique en arabe

6.2.1. Mesure de similarité sémantique locale basée sur « DiLAC »

6.2.1.1. DiLAC

DiLAC est une nouvelle version électronique du « Dictionnaire de la langue arabe contemporaine »[مختا08], et nous sommes les premiers, à notre connaissance, qui l'utiliser pour mesurer la similarité sémantique et pour la désambiguïsation sémantique des mots arabes. DiLAC est une vaste base de données lexicale en arabe ; il comporte une quantité énorme d'information par entrée (voir chapitre 3), ce qui le serait potentiellement capable de faire plus de distinctions valides entre les sens d'un mot polysémique, qu'un dictionnaire disposant d'un matériau moins riche [Lesk86]. Par exemple, Lesk rapporte que pour le mot « galley » dans le contexte de la phrase « stoke the stove in the galley », le dictionnaire « OALDCE » ne fournit aucune intersection de sens entre « galley » et « stove » et produit une désambiguïsation incorrecte. En revanche, le ditionnaire « OED », incluant dans la définition de sens (2) de « galley » des mots tels que : « stove », « cook », « cooking-room » et « pot », mène à un bon résultat.

DiLAC, dans son version originale, est un modèle pour décrire les concepts et les relations entre eux, dont les noms (singulier, duel, pluriel), les verbes, les mots fonctionnels sont organisés en un ensemble d'entrées, et cette organisation nous permet de proposer une structure hiérarchique, en utilisant un fichier basé sur XML conforme à LMF[35] (Lexical Markup Framework), ce qui en fait un outil utile pour la linguistique informatique et le traitement du langage naturel.

6.2.1.1.1. Les Entrées de DiLAC

Dictionnaire monolingue arabe réalisé par Ahmed Mukhtar Abdul Hamid Omar avec l'aide d'un groupe de travail. Le dictionnaire contemporain contient 5778 racines, 32300 entrées lexicales (10475 verbes, 21457 noms et 368 particules), 32300 entrées et 43384 exemples supplémentaires, 63019 significations et 17883 expressions contextuelles. Comme il contient les domaines des concepts, par exemple : sports et éducation physique ; culture et arts ; agriculture ; médecine ; géographie ; etc. Ainsi que de nombreuses autres informations lexicales, syntaxiques et sémantiques ce qui en fait une ressource lexicale intéressante pour

[35] Lexical Markup Framework (Cadre de balisage lexical, en français) est le standard de l'Organisation internationale de normalisation (au sein de l'ISO/TC37) pour les lexiques du traitement automatique des langues. L'objectif est de fournir un modèle commun pour la création et l'utilisation des ressources langagières, de gérer l'échange des données entre ces ressources et de permettre la fusion d'un grand nombre de ressources électroniques afin de constituer un vaste réseau de descriptions linguistiques.

WSD et les systèmes de recherche d'information sémantiques en arabe, par rapport à AWN (Arabic WordNet), la plus utilisée dans ce domaine de recherche, et qu'elle contient, dans sa première version, 9698 concepts, correspondant à 21813 mots, et 6 types de relations différents, et dans une version ultérieure, a également été publiée et contient 11269 synsets, ce qui correspond à 23841 mots et 22 types de liens. Les synsets AWN appartiennent à l'une des cinq parties du discours : nom (6438), verbe (2536), adjectif (456), satellite adjectif (158) et adverbe (110) [CSBA13].

6.2.1.1.2. Structure de DiLAC

Nous avons effectué plusieurs opérations sur « le dictionnaire de langue arabe contemporaine » pour représenter dans une nouvelle version électronique robuste et exploitable par le module de mesure de similarité et le module WSD, et comme première pas, nous avons récupéré et étudié la version texte de dictionnaire pour extraire les entrées les plus importantes de dictionnaire. Après une analyse détaillée du contenu de dictionnaire, nous avons conçu ses entrées comme suit :

1. *Dict* : est la racine du dictionnaire ;
2. *wordEntry* : est l'entrée pour chaque mot dans le dictionnaire avec le paramètre « *idWordEntry* » qui présente l'identifiant de l'entrée du mot dans le dictionnaire ;
3. *word* : est le mot avec un paramètre « *idWord* » qui présente l'identifiant du mot ;
4. *rootLetters* : sont les lettres du lemme du mot ;
5. *property* : est la partie de discours avec le paramètre « *POS* » (Part of speech) ;
6. *explanation* : est explication d'une signification particulière du mot avec deux paramètres : « *idx* » présente l'identifiant de chaque explication ; et « *defx* » présente la définition de ce mot, en indiquant son domaine scientifique, s'il existe, dans le paramètre « *filed* » ;
7. *example* : est explication d'une signification particulière du mot, par l'utilisation d'exemple, avec deux paramètres : « *ide* » présente l'identifiant de chaque exemple ; et « *defe* » présente l' explication de cet exemple, en indiquant son domaine scientifique, s'il existe, dans le paramètre « *filed* » ;
8. *addExampeles* : sont les exemples supplémentaires, s'ils existent, sur chaque signification avec deux paramètres : « *aeId* » présente l'identifiant de chaque exemple supplémentaire ; et « *defe* » présente la phrase de cet exemple.

La Figure 6.2 illustre un aperçu de la structure résultant de ce processus de conception.

Figure 6.2: Un aperçu sur DiLAC

6.2.1.1.3. Le dictionnaire DiLAC-Lesk

Le module de mesure de similarité et le module WSD nécessitent, pour leurs exécutions, un format de dictionnaire spécifique (Figure 6.3), dont ses définitions sont codées en entiers et triés par ordre croissant, à la place des chaines de caractères, pour faciliter et accélérer les tâches de comparaison et de recherche. En revanche, pour générer un DiLAC-Lesk à partir de la version originale (DiLAC), nous avons appliqué les mêmes traitements que nous suivions dans l'approche d'indexation [DiBe12] avec quelques modifications (Figure 6.4) :

```
<ids>آبنوس%1:05:00::</ids>
<def>8 22 112 428 528 705 1216 1228 1538 2450 2518 7518 9842 31822 45730
</def>
```

Figure 6.3: Un exemple sur le format de dictionnaire DiLAC-Lesk

1. Nous devons convertir le contenu de DiLAC en codage UTF-16, car il permet la représentation de lettres et de symboles dans un large éventail de langues, y compris l'arabe.
2. Nous appliquons les mêmes étapes de normalisation, suppression des mots vides et lemmatisation vues dans le chapitre précédent.

3. Nous appliquons la méthode de lemmatisation hybride [DiBe12] qui donne une meilleure performance dans la recherche d'information.
4. Nous ajoutons une fonction qui générer des identifiant de type entier pour chaque mot apparais dans DiLAC. On affecte un entier à chacun des mots.
5. Ensuite, nous regroupons les ensembles d'entiers qui codent les mots, par définitions et ces ensembles sont triés par ordre croissant.

DiLAC-Lesk_Generate function pseudo code

```
Input: Dictionary DiLAC
Output: Dictionary DiLAC_Lesk_format
begin
      For each entry Ei in Dict loop
            Ei = "";
            For each word wi in Ei loop

                  Encoding ();
                  Normalize ();
                  Removing_stop_words ();
                  Stemming ();
                  If (stem doesn't have an id) then
                        Stem_id = Generate_id(stem)
                  End
                  NEi = NEi + Stem_id;
            End Loop
      End loop.
      Sorted in ascending order (NEi)
      Add NEi to DiLAC_Lesk_format;
end.
```

Figure 6.4: Algorithme de génération de DiLAC-Lesk

6.2.1.2. Mesures de similarité sémantique

La similarité basée sur la connaissance est l'une des mesures de similarité sémantique qui repose sur l'identification du degré de similarité entre les mots à l'aide d'informations dérivées de réseaux sémantiques [MiCS06].

Les mesures de similarité basées sur la connaissance peuvent être divisées en deux groupes : les mesures de similarité sémantique et les mesures de relation sémantique. Les concepts sémantiquement similaires sont réputés être liés sur la base de leur ressemblance. La relation sémantique, en revanche, est une notion plus générale de la relation, qui n'est pas spécifiquement liée à la forme du concept. En d'autres termes, la similitude sémantique est une sorte de parenté entre deux mots, elle recouvre un éventail plus large de relations entre

concepts, qui inclut des relations de similitude supplémentaires telles que Est-une-sorte-de ; est-un-exemple-spécifique-de ; est-une-partie-de, etc. [PaBP03].

Il y a six mesures de similarité sémantique ; trois d'entre elles sont basées sur le contenu de l'information : Resnik (*res*) [Resn95], Lin (lin) [Lin98b] et Jiang & Conrath (*jcn*) [JiCo97]. Les trois autres mesures sont basées sur la longueur du chemin : Leacock & Chodorow (*lch*) [Chod98], Wu & Palmer (*wup*) [WuPa94] et Path Length (*chemin*).

En outre, il existe trois mesures de relation sémantique : Hirst & StOnge (hso) [HiSt98], Lesk [BaPe02] et paires de vecteurs (vecteur) [Patw03].

La mesure Lesk fonctionne en trouvant des superpositions dans les gloses des deux définitions. Le score de similarité est la somme des carrés des longueurs de superpositions.

Dans ce travail, nous sélections deux mesures, pour comparer leurs résultats avec la nôtre qui basée sur DiLAC. La première mesure sélectionné est Wup, car elle a donné les meilleurs résultats, et la deuxième mesure sélectionné est la mesure de référence du système AWSS (Arabic Word Semantic Similarity), car il s'agit de la première mesure de similarité sémantique arabe et permet de comparer ses résultats sur un jeu de données de référence (Benchmark Dataset) arabe avec les résultats de la nôtre. AWSS proposé par Almarsoomi et al. [AOBC13] pour calculer la similarité entre les concepts utilisant des sources d'information extraites de AWN (Arabic WordNet), ces informations sont la longueur et la profondeur, ils ont également utilisé un jeu de données de référence arabes [FJZK12] développé pour évaluer une mesure AWSS en calculant la similarité de mots sur un ensemble de mots arabes avec des jugements humains, et ils ont constaté que l'évaluation expérimentale indique que la mesure arabe se comporte bien. Il a atteint une valeur de corrélation de 0,894 par rapport à la valeur moyenne des participants humains de 0,893 sur le jeu de données de référence [AOBC13].

6.2.1.2.1. Jeu de données de référence arabe

Dans cette étude, le jeu de données de référence arabe utilisé est appelé AWSS, et cet ensemble de données a été créé par Fazza et al [FJZK12], AWSS utilise les mêmes procédures qui ont été suivies lors de la création des versions anglaises les plus connus ; Rubenstein & Goodenough [RuGo65] et Miller & Charles [MiCh91]. À notre connaissance, il n'existe pas d'ensembles de données de référence arabes pour la similarité sémantique, à l'exception de AWSS.

Le jeu de données de référence AWSS a été préparé principalement en deux étapes : premièrement, déterminer le jeu de paires de mots arabes ; ensuite, spécifier le taux de

similarité humaine pour les paires de mots, dont ils ont créé une liste de paires de mots arabes contenant 70 pairs.

Ils ont créé et utilisé 27 catégories arabes et pour sélectionner les paires de mots de stimulation arabe et pour promouvoir la meilleure représentation sémantique possible, cinq autres catégories ont été ajoutées pour élargir 22 catégories à 27 catégories. Les paires de noms anglais de la liste Rubenstein & Goodenough ont été traduites et vérifiées pour créer les 22 catégories arabes. Une fois que les 22 catégories ont été spécifiées, 5 nouvelles catégories ont été ajoutées, correspondant au style de vie arabe, après cela, les deux premiers noms de chaque catégorie sont sélectionnés pour générer 56 mots arabes.

Les 56 paires de noms ont été divisées en deux colonnes, 28 noms dans chaque colonne. Et un échantillon de 22 locuteurs natifs de 5 pays arabes différents a été choisi pour générer deux paires de noms arabes, allant d'une similarité de sens élevée à une similarité moyenne et à une similarité faible. Il a été demandé aux participants d'écrire 28 paires de noms arabes qui présentent une similarité élevée dans la liste en sélectionnant un nom de la colonne A et un autre de la colonne B, et d'écrire 32 paires présentant une similarité moyenne de la même manière, et 13 paires de noms arabes de faible similarité ont été sélectionnées de manière aléatoire par Fazza et al. La liste finale contenait 70 paires de noms arabes qui couvraient une similarité élevée à faible, cette liste a été appelée jeu de données de référence AWSS. Le Tableau 6.1 présente la liste AWSS.

Soixante autres participants originaires de différents pays arabes qui n'avaient pas participé à la création de paires de mots arabes ont été invités à classer l'ensemble de ces 70 paires de mots arabes, ces participants ont été priés d'évaluer chaque paire de mots en fonction de leur similarité sémantique, et cette évaluation est une note de 0,0 à 4,0.

	Paires de mots en anglais		*Paires de mots en arabe*		*Evalu. Huma.*		*Paires de mots en anglais*		*Paires de mots en arabe*		*Evalu. Huma.*
1	Coast	Endorsement	تصديق	ساحل	0.03	*36*	Slave	Lad	فتى	عبد	1.77
2	Noon	String	خيط	ظهر	0.03	*37*	Journey	Bus	باص	رحلة	1.83
3	Cushion	Diamond	الماس	مسند	0.06	*38*	Girl	Odalisque	جارية	فتاة	1.96
4	Gem	Pillow	مخدة	جوهرة	0.07	*39*	Feast	Fasting	صيام	عيد	1.96
5	Stove	Walk	مشي	موقد	0.07	*40*	Coach	Means	وسيلة	حافلة	2.07
6	Cord	Midday	ظهيرة	حبل	0.08	*41*	Brother	Lad	فتى	أخ	2.15
7	Signature	String	خيط	توقيع	0.08	*42*	Sage	Sheikh	شيخ	حكيم	2.26
8	Boy	Endorsement	تصديق	صبي	0.12	*43*	Girl	Sister	اخت	فتاة	2.38
9	Boy	Midday	ظهيرة	صبي	0.16	*44*	Hill	Mountain	جبل	تل	2.6
10	Slave	Vegetable	خضار	عبد	0.16	*45*	Hen	Pigeon	حمامة	دجاجة	2.61
11	Smile	Village	قرية	ابتسامة	0.18	*46*	Master	Sheikh	شيخ	سيد	2.66
12	Smile	Pigeon	حمامة	ابتسامة	0.2	*47*	Food	Vegetable	خضار	طعام	2.78
13	Wizard	Infirmary	مشفى	ساحر	0.22	*48*	Slave	Odalisque	جارية	عبد	2.84
14	Noon	Fasting	صيام	ظهر	0.29	*49*	Run	Walk	مشي	جري	3.01
15	Hill	Pigeon	حمامة	تل	0.33	*50*	Brother	Sister	أخت	أخ	3.08

16	Countryside	Laugh	ضحك	ريف	0.34	*51*	Cord	String	خيط	حبل	3.09
17	Glass	Diamond	الماس	كأس	0.36	*52*	Forest	Woodland	أحراش	غابة	3.14
18	Glass	Fasting	صيام	كأس	0.38	*53*	Sage	Thinker	مفكر	حكيم	3.3
19	Cord	Mountain	جبل	حبل	0.54	*54*	Gem	Diamond	الماس	جوهرة	3.38
20	Hospital	Grave	قبر	مستشفى	0.83	*55*	Cushion	Pillow	مخدة	مسند	3.38
21	Forest	Shore	شاطئ	غابة	0.86	*56*	Journey	Travel	سفر	رحلة	3.39
22	Gem	Young woman	شابة	جوهرة	0.87	*57*	Countrysid	Village	قرية	ريف	3.41
23	sepulcher	Sheikh	شيخ	ضريح	0.89	*58*	Smile	Laugh	ضحك	ابتسامة	3.48
24	Tool	Pillow	مخدة	أداة	0.99	*59*	Stove	Oven	فرن	موقد	3.55
25	Coast	Mountain	جبل	ساحل	1.06	*60*	Coast	Shore	شاطئ	ساحل	3.56
26	Run	Shore	شاطئ	جري	1.13	*61*	Signature	Endorsement	تصديق	توقيع	3.58
27	Hill	Woodland	أحراش	تل	1.19	*62*	Tool	Means	وسيلة	أداة	3.68
28	Countryside	Vegetable	خضار	ريف	1.24	*63*	Noon	Midday	ظهيرة	ظهر	3.7
29	Tumbler	Tool	قدح	أداة	1.32	*64*	Boy	Lad	فتى	صبي	3.71
30	Master	Thinker	مفكر	سيد	1.36	*65*	Girl	Young Woman	شابة	فتاة	3.74
31	Feast	Laugh	ضحك	عيد	1.36	*66*	Sepulcher	Grave	قبر	ضريح	3.75
32	Hen	Oven	فرن	دجاجة	1.44	*67*	Wizard	Magician	مشعوذ	ساحر	3.76
33	Journey	Shore	شاطئ	رحلة	1.47	*68*	Coach	Bus	باص	حافلة	3.8
34	Coach	Travel	سفر	حافلة	1.6	*69*	Glass	Tumbler	قدح	كأس	3.82
35	Food	Oven	فرن	طعام	1.76	*70*	Hospital	Infirmary	مشفى	مستشفى	3.91

Tableau 6.1: Jeu de données de référence AWSS

Dans ce travail, le jeu de données de référence AWSS a été choisi pour les raisons suivantes : premièrement, on n'a pas d'autre choix, car comme nous avons déjà mentionné, AWSS est le seul jeu de données de référence pour la langue arabe, deuxièmement, les paires de mots arabes ont été créées avec soin.

L'absence de certains mots dans AWN d'une part, et d'autre part pour que nous puissions comparer les résultats de [MoMo17] avec le nôtre, nous été obligé de prendre 40 paires de mots arabes : 12 paires de mots de faible similarité, 13 paires de mots de moyenne similarité et 15 paires de mots de forte similarité. En outre, nous convertissons toutes les valeurs de AWSS de l'intervalle [0.0-0.4] à [0.0-1.0] en divisant par quatre.

6.2.1.2.2. Expérimentation des mesures de similarité sur les Dictionnaire

Dans cette section, nous étudierons l'efficacité de DiLAC à la désambiguïsation sémantique des mots arabes à travers d'utiliser les mesures de similarité sémantique. Les résultats de cette étude fourniront aux chercheurs en traitement de la langue arabe une autre ressource lexicale robuste pouvant être utilisées dans leurs futures recherches.

L'étude expérimentale présentée dans cette section est organisée comme suit : sélection des outils d'application des deux mesures de similarité sémantique sur AWN, application des mesures de similarité sémantique traditionnelles à l'aide de l'outil sélectionné, extraction et analyse des résultats de la mise en œuvre des mesures, évaluation finale des résultats basés sur le MSE et la corrélation.

6.2.1.2.3. L'outils de mesure de similarité sur DiLAC

Nous développons un outil de mesure de similarité sémantique sur DiLAC, en utilisant le langage de programmation Java et en se basant sur l'algorithme de Lesk, dont il consiste simplement à compter les mots communs entre les définitions et les exemples d'usage des sens de deux paire de mot. On choisit alors le sens caractérisé par le plus grand nombre de superpositions (Overlap).

Dans cette méthode, on a également s'appuie sur : des connaissances de nature syntaxique ; et le domaine (le nom de champs scientifique) de chaque mot (sport, politique, religion, etc.) offerts par DiLAC.

Plus formellement, soit la paire de mots (m_1, m_2) à calculer sa mesure de similarité. Si on considère ses définitions formé par les deux phrase $p_1$et p_2, comportant les mots $m_i, i = 1, \dots, k$, et $m_2, j = 1, \dots, l$, alors le score de mesure de similarité est déterminé par la formule :

$$similarite\_semantique\ (m_1, m_2) = \sum_{i=1}^{k} poids(m_i) \qquad \text{(Eq.6.1)}$$

Avec

$$poids(m_2) = \begin{cases} -log\big(p(m_i)\big), & si\ m_i \in D(m_1) \\ 0, & si\ non \end{cases} \qquad \text{(Eq.6.2)}$$

Où

$D(m_2)$: représente l'ensemble de mots composant la définition du m_2 (+ exemples d'usage) et p(m_2) est la fraction des définitions et des exemples du dictionnaire qui contient le mot m_2 :

$$p(m_2) = \frac{nbre\ de\ définition\ et\ d'exemples\ du\ DiLAC\ qui\ contient\ m_2}{nbre\ total\ de\ définition\ et\ d'exemples\ du\ dictionnaire} \qquad \text{(Eq.6.3)}$$

6.2.1.2.4. Calcul de la similarité sémantique et comparaison

Le but de cette section est de présenter d'autres approches basées sur d'autres algorithmes ou sur l'exploitation de WordNet et Arabic WordNet, afin de mieux retracer l'encadrement comparatif de notre recherche.

Dans [MoMo17], l'API Java AWN et WS4J sont utilisés, dont L'API Java AWN contient des implémentations de quatre mesures de similarité sémantiques, Wup, LCH, LI et Path (chemin), en outre, il fournit des sources d'informations telles que le nombre d'hyponymes pour les concepts, la profondeur des concepts dans la taxonomie et la longueur de chemin d'accès entre les concepts. Les quatre mesures mentionnées sont appliquées, ainsi qu'une mesure supplémentaire appelée Resnik qui repose sur les informations fournies par l'outil. WS4J est le deuxième outil utilisé pour calculer la similarité sémantique sur les paires de noms anglais. Cet outil peut calculer le score de similarité en utilisant huit mesures sur WN, outil facile à utiliser en ligne.

Dans cette section, les mesures de similarité sémantique seront appliquées à l'aide de l'API AWN java sur 40 paires de noms arabes sélectionnées à partir du jeu de données AWSS. Le résultat de toutes les mesures sera décrit, analysé et comparé aux évaluations humaines.

		Paires de mots en anglais		Paires de mots en arabes		Eval. Hum.	Wup			WASS			Lesk-ar		
							Sco.	Err.	Err. Quad.	Sco.	Err.	Err. Quad.	Sco	Err.	Err. Quad.
1	Similarité faible	Coast	Endorsement	تصديق	ساحل	0.01	0	0.01	0.0001	0	0.01	0.0001	0.03	-0.02	0.0001
2		Noon	String	خيط	ظهر	0.01	0	0.01	0.0001	0.17	-0.16	0.0256	0.03	-0.02	0.0001
3		Stove	Walk	مشي	موقد	0.02							0.15	-0.13	0.0169
4		Cord	Midday	ظهيرة	حبل	0.02	0	0.02	0.0004	0	0.02	0.0004	0	0.02	0.0004
5		Signature	String	خيط	توقيع	0.02	0	0.02	0.0004	0	0.02	0.0004	0.05	-0.03	0.0011
6		Boy	Endorsement	تصديق	صبي	0.03	0	0.03	0.0009	0	0.03	0.0009	0.03	-0.01	0
7		Boy	Midday	ظهيرة	صبي	0.04	0	0.04	0.0016	0	0.04	0.0016	0	0.04	0.0016
8		Smile	Village	قرية	ابتسامة	0.05	0	0.05	0.0025	0	0.05	0.0025	0.02	0.02	0.0006
9		Noon	Fasting	صيام	ظهر	0.07	0	0.07	0.0049	0	0.07	0.0049	0.26	-0.19	0.0397
10		Glass	Diamond	الماس	كأس	0.09	0.12	-0.03	0.0009	0.05	0.04	0.0016	0.06	0.03	0.0009
11		sepulcher	Sheikh	شيخ	ضريح	0.22	0.18	0.04	0.0016	0.06	0.16	0.0256	0.2	0.02	0.0004
12		Countryside	Vegetable	خضار	ريف	0.31	0.18	0.13	0.0169	0.45	-0.14	0.0196	0.5	-0.19	0.0361
13	Similarité moyenne	Tumbler	Tool	قدح	أداة	0.33	0.5	-0.17	0.0289	0.54	-0.21	0.0441	0.5	-0.17	0.0289
14		Laugh	Feast	عيد	ضحك	0.34	0.15	0.19	0.0361	0.66	-0.32	0.1024	0.36	-0.02	0.0004
15		Girl	Odalisque	جارية	فتاة	0.49	0.54	-0.05	0.0025	0.73	-0.24	0.0576	0.3	0.19	0.0361
16		Feast	Fasting	صيام	عيد	0.49	0.18	0.31	0.0961	0.17	0.32	0.1024	0.4	0.09	0.0081
17		Coach	Means	وسيلة	حافلة	0.52	0.66	-0.14	0.0196	0.38	0.14	0.0196	0.49	0.03	0.0009
18		Sage	Sheikh	شيخ	حكيم	0.57	0.46	0.11	0.0121	0.67	-0.1	0.01	0.2	0.37	0.1369
19		Girl	Sister	اخت	فتاة	0.6	0.54	0.06	0.0036	0.37	0.23	0.0529	0.5	0.1	0.01
20		Hen	Pigeon	حمامة	دجاجة	0.65	0.78	-0.13	0.0169	0.89	-0.24	0.0576	0.6	0.05	0.0025
21		Hill	Mountain	جبل	تل	0.65							0.58	0.06	0.0044
22		Master	Sheikh	شيخ	سيد	0.67	0.5	0.17	0.0289	0.67	0	0	0.55	0.12	0.0144
23		Food	Vegetable	خضار	طعام	0.69	0.4	0.29	0.0841	0.53	0.16	0.0256	0.9	-0.21	0.0441
24		Slave	Odalisque	جارية	عبد	0.71	0.66	0.05	0.0025	0.93	-0.22	0.0484	0.67	0.04	0.0016
25		Run	Walk	مشي	جري	0.75	0.83	-0.08	0.0064	0.6	0.15	0.0225	0.9	-0.15	0.0225
26	Similarité élevé	Cord	String	خيط	حبل	0.77	0.66	0.11	0.0121	0.7	0.07	0.0049	0.57	0.2	0.04
27		Forest	Woodland	أحراش	غابة	0.79	0.88	-0.09	0.0081	0.82	-0.03	0.0009	0.7	0.09	0.0081
28		Sage	Thinker	مفكر	حكيم	0.83	0.8	0.03	0.0009	0.75	0.08	0.0064	0.8	0.03	0.0009
29		Journey	Travel	سفر	رحلة	0.85	0.9	-0.05	0.0025	0.87	-0.02	0.0004	0.9	-0.05	0.0025
30		Gem	Diamond	الماس	جوهرة	0.85	0.83	0.02	0.0004	0.89	-0.04	0.0016	0.9	-0.05	0.0025
31		Countryside	Village	قرية	ريف	0.85	0.8	0.05	0.0025	0.82	0.03	0.0009	0.92	-0.07	0.0049
32		Cushion	Pillow	مخدة	مسند	0.85	0.57	0.28	0.0784	0.82	0.03	0.0009	0.6	0.25	0.0625
33		Smile	Laugh	ضحك	ابتسامة	0.87	0.62	0.25	0.0625	0.29	0.58	0.3364	0.8	0.07	0.0049
34		Signature	Endorsement	تصديق	توقيع	0.9	0.8	0.1	0.01	0.93	-0.03	0.0009	0.55	0.34	0.1186
35		Tool	Means	وسيلة	أداة	0.92	0.76	0.16	0.0256	0.93	-0.01	0.0001	0.69	0.23	0.0509
36		Sepulcher	Grave	قبر	ضريح	0.94	1	-0.06	0.0036	0.82	0.12	0.0144	0.9	0.04	0.0016
37		Boy	Lad	فتى	صبي	0.93	0.88	0.05	0.0025	0.95	-0.02	0.0004	0.8	0.13	0.0169
38		Wizard	Magician	مشعوذ	ساحر	0.94							0.9	0.04	0.0016
39		Coach	Bus	باص	حافلة	0.95	1	-0.05	0.0025	0.94	0.01	0.0001	0.97	-0.02	0.0004
40		Glass	Tumbler	قدح	كأس	0.95	0.77	0.18	0.0324	0.89	0.06	0.0036	0.65	0.29	0.0858
MSE							0.016540541			0.026978378			0.020308627		
MSE faible							0.002754545			0.007563636			0.008284841		
MSE moyenne							0.028141667			0.045258333			0.023911111		
MSE élevé							0.017428571			0.026564286			0.026805504		
Corrélation							0.941168501			0.889771136			0.916607931		
Corrélation faible							0.918748881			0.756095628			0.840386909		
Corrélation moyenne							0.598689836			0.293223608			0.657098832		
Corrélation élevé							0.41198363			0.350032563			0.31006654		

Tableau 6.2: Résultats de l'application des mesures Wup, AWSS sur AWN et Lesk-Ar sur DiLAC

6.2.1.2.5. Analyse des résultats

Après avoir calculé le score de similarité pour toutes les paires de mots arabes et récupéré les résultats de [MoMo17], l'étape suivante consistait à représenter les valeurs de similarité de : Lesk_DiLAC, Wup et AWSS, puis à étudier les performances de toutes les mesures. Par conséquent, nous avons écrit les résultats dans le Tableau 6.2. Le processus d'évaluation dans ce travail a été réalisé en fonction de deux facteurs, à savoir : la corrélation entre le score de mesure de similarité et la notation humaine ; et l'erreur quadratique moyenne[36] (MSE : Mean Squared Error) des résultats des mesures. Les résultats de l'application des mesures sur les 40 paires de mots arabes ont été comparés pour étudier les différences entre AWN et DiLAC. La colonne « Éval. Hum. » qui contient le score de l'évaluation humain, considère comme une référence pour calculer le taux d'erreur de la mesure de similarité sémantique, ce score est utilisé pour la comparaison avec les résultats obtenus par Wup, AWSS et Lesk-ar. Les deux colonnes (Err., Err._Quad.) contiennent respectivement : l'erreur, qui correspond à la différence entre le score de similarité calculé par les différentes approches et l'évaluation humaine ; et l'erreur carrée permettant de calculer l'erreur quadratique moyenne. Les paires de mots sont divisées en trois catégories : similarité faible, similarité moyenne et similarité élevé.

La colonne Wup montre que la mesure de Wup a obtenu des bons résultats : le MSE est (0,016475), la valeur MSE de la catégorie « similarité élevé » est (0,01740) et les valeurs MSE des catégorie « similarité moyen » et « similarité faible » sont respectivement (0,0027) et (0.028). Ces résultats indiquent une meilleure performance pour Wup avec une similarité élevée. La mesure Wup a obtenu une valeur élevée du coefficient de corrélation (0,94) par rapport à l'évaluation humaine, ce qui signifie que la mesure Wup a une bonne relation linéaire avec l'évaluation humaine. La Figure 6.5 montre la corrélation entre les évaluations humaines et les scores de mesure de Wup.

[36] L'erreur quadratique moyenne d'un estimateur θ d'un paramètre θ de dimension 1 est une mesure caractérisant la « précision » de cet estimateur. Le fait que cette mesure soit presque toujours strictement positive (et non nulle) est dû à un caractère aléatoire ou au fait que l'estimateur ne prend pas en compte les informations susceptibles de produire une estimation plus précise.

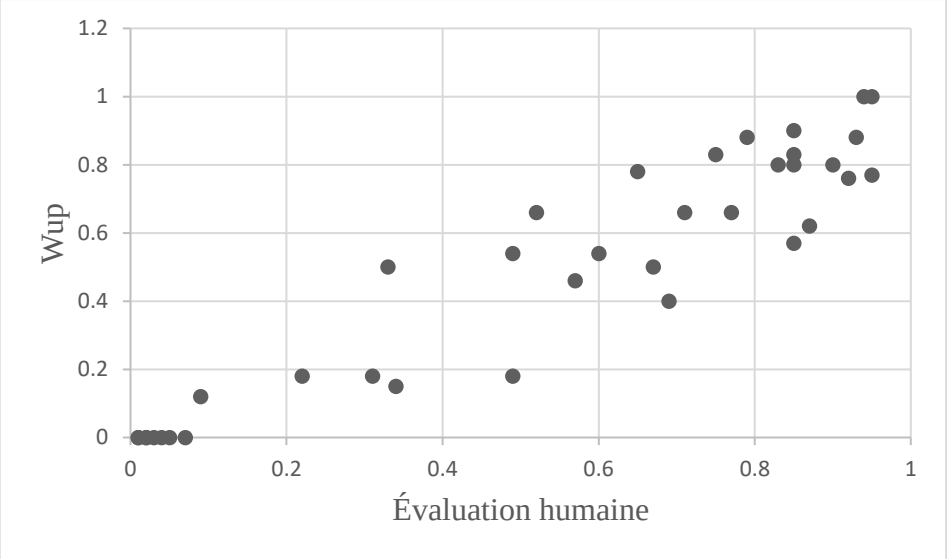

Figure 6.5: La corrélation entre Wup et l'évaluation humaine

La colonne AWSS montre que cette mesure permet d'obtenir un mauvais score MSE (0.026979) par rapport aux deux autres méthodes. La corrélation de notation humaine avec la méthode AWSS (0.889771) est très proche de la corrélation de scores humains. La Figure 6.6 montre la corrélation entre les scores de la mesure AWSS et les évaluations humaines.

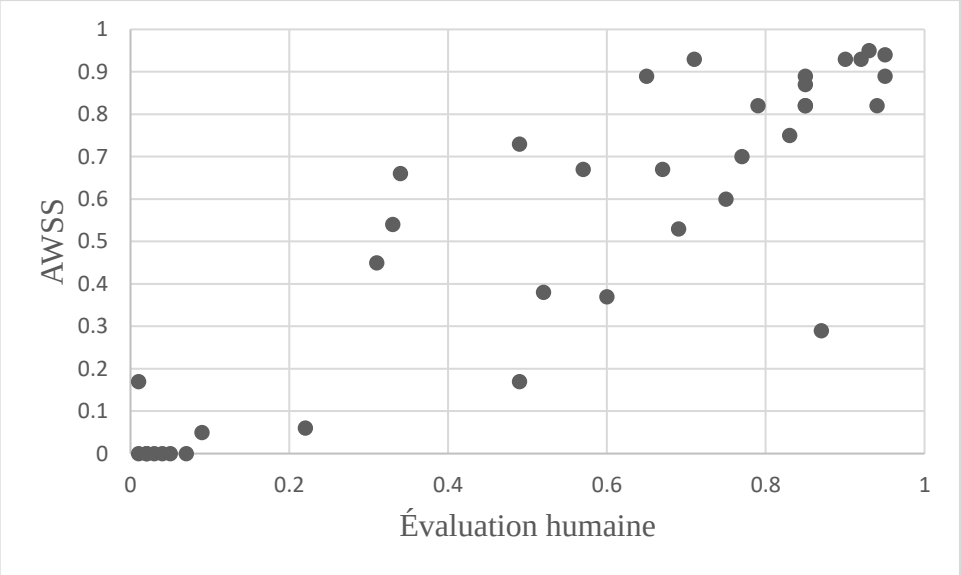

Figure 6.6: La corrélation entre AWSS et l'évaluation humaine

La dernière mesure appliquée est la nouvelle mesure Lesk-ar, comme indiqué dans le Tableau 6.2, la valeur MSE (0.020301) de cette mesure est meilleur que la mesure AWSS et elle est très proche de la valeur MSE de la mesure Wup, ainsi que la valeur MSE (0.023911) de la catégorie « similarité élevé » indique la puissance de cette mesure. La Figure 6.7 montre

la corrélation entre la mesure de Lesk-ar et l'évaluations humaine. Cette mesure présente un score très proche de la mesure Wup et montre une forte corrélation (0.9166).

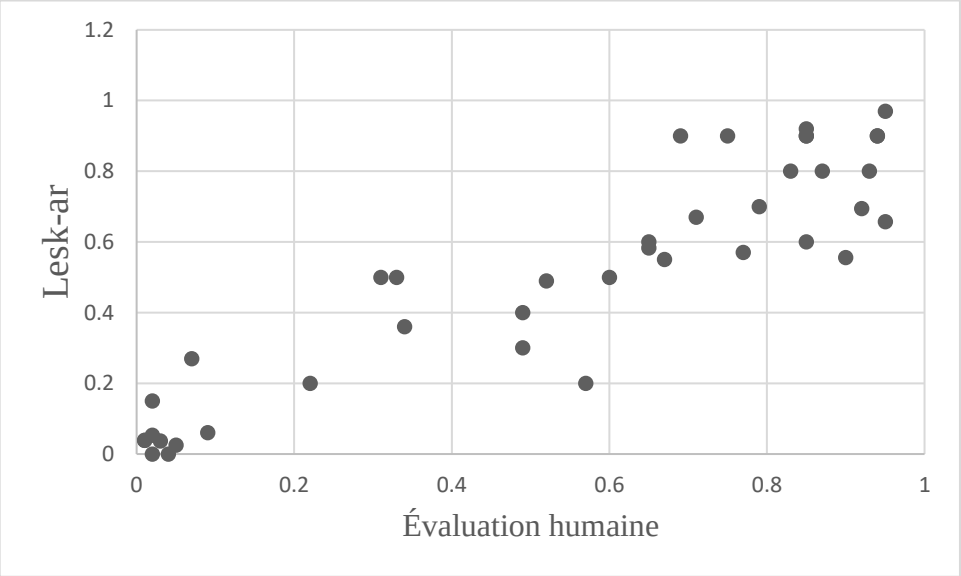

Figure 6.7: La corrélation entre Lesk-ar et l'évaluation humaine

Nous avons étudié, dans cette section la possibilité d'appliquer une mesure de similarité sémantiques traditionnelle sur DiLAC, et cette mesure a été appliquée à l'aide d'un jeu de données de référence arabe appelé AWSS. Le DiLAC est une ressource lexicale arabe riche fournit des définitions et des exemples d'usages et d'autre information syntaxiques et lexicales (voir chapitre 3) utilisé par les mesures basées basée sur le recouvrement de traits entre deux mots, et permettant de calculer le score de similarité entre des paires de mots arabes.

Les mesures basées basée sur le recouvrement de traits entre deux mots utilisent les informations fournis par DiLAC pour calculer le score de similarité entre des paires de mots arabes. Par exemple, la mesure de Lesk, qui est une mesure à base de traits, a donné des très bons résultats, en exploitant DiLAC, cependant, elle ne s'applique pas à AWN.

6.2.2. Algorithme global : approche exhaustive pour la désambiguïsation sémantique des mots arabes

L'algorithme que nous avons implémenté est une variante simplifiée s'appuient principalement sur la méthode originelle de Lesk [Lesk86], adaptée aux caractéristiques constructives de DiLAC.

Les entités descriptives de sens utilisées dans la désambiguïsation, sont d'un côté, les définitions et les exemples d'usages, et d'un autre, les relations de type synonymie et hyperonymie, ainsi qu'une combinaison des deux (définitions + exemples et relations).

Chaque entité descriptive est représentée par un sac de mots, c.-à-d. une collection de mots dont l'ordre et la dépendance sont ignorés. Les implémentations permettent le choix du type de description, de la fenêtre de contexte et supposent que les sens candidats du mot à désambiguïser ont été préalablement ordonnés, en ordre décroissant de leur fréquence d'usage. L'ordonnancement des sens est effectué pendant la phase de lemmatisation, à partir de l'information disponible dans DiLAC.

En ce qui concerne la dépendance entre les sens déjà assignés et ceux en cours d'assignation, on choisit le meilleur candidat pour un mot cible donné, sans que ce choix influence ultérieurement la désambiguïsation du mot suivant.

La Figure 6.8 représente le mot à désambiguïser tw, dans son contexte $C(tw)$ qui contient entre autres mots le mot w. A chaque sens S_j de tw correspond une définition $D(S_j)$ dans le dictionnaire. Le mot w est représenté dans le dictionnaire par la réunion des définitions de ses sens, $E(w)$.

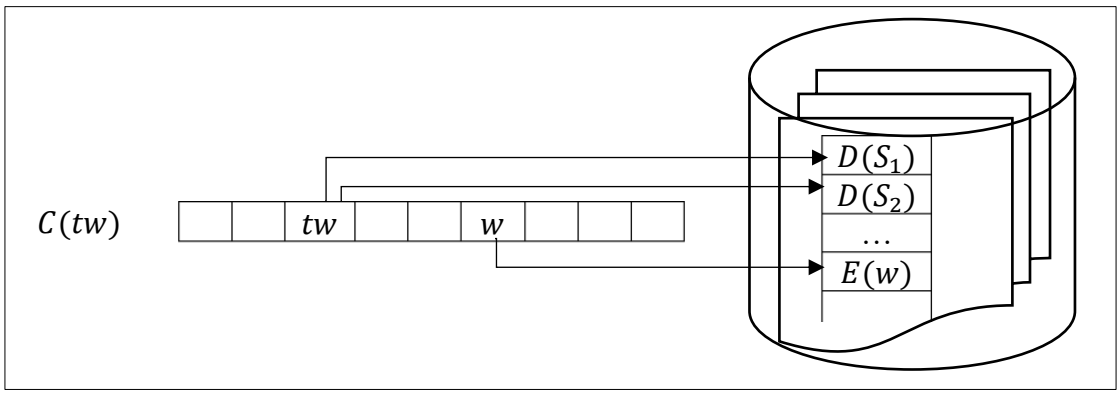

Figure 6.8: Schéma de l'algorithme de Lesk de base

La variante simplifiée est similaire à celle de base, la seule différence est que pour le calcul du score on compte les superpositions entre l'entité descriptive du sens candidat $D(S_j)$ et les mots du contexte w (et non plus leur définitions). Soit $C(tw)$ la fenêtre de contexte formée par le sac de mots w, en forme de base, alors la représentation en pseudo-code de cette variante est comme suite :

```
Lesk_WSD function pseudo code
begin
      For each w wi in loop
            best_score = 0;
            best_candidate = s1
            sup = 0;
            define C(w) // define the context of m
            For each candidate sense sj of w loop
                  extract the definition D from DiLAC;
                  Calculate overlap number // see Eq.6.4
                  If (best_score < overlap) then
                        best_score = overlap;
                        best_candidate = sj
                  End
            End Loop
      End loop.
      w = best_candidate;
end.
```

Figure 6.9: Algorithme de Lesk simplifié

La définition $D(S_j)$ et la fenêtre de contexte $C(w)$ formée par le sacs de lemmes (mots après lemmatisation : voir chapitre 4). Le nombre de superpositions entre la définition d'un sens candidat et la définition d'un mot du contexte est calculé comme le nombre d'éléments de l'intersection des deux sacs de mots. Pour chaque mot à désambiguïser, l'algorithme assigne initialement, comme meilleur candidat, le sens le plus fréquent (S_1, le premier dans l'ordre des sens). Un autre sens est choisi si et seulement si son score est supérieur à celui du meilleur candidat courant.

$$overlap = \left|D(S_j) \cap C(w)\right| \qquad \text{(Eq.6.4)}$$

Dans notre implémentation, les entités descriptives $D(S_j)$, $C(w)$ désignent des définitions, exemples d'usage et le domaine (champs scientifique de mots : voir chapitre3), éléments extraits de DiLAC.

En outre, Nous avons normalisé par $\log_2$ de la taille de $D(S_j)$, et la raison de ces normalisations réside dans le fait que les descriptions trop longues tendent à dominer les plus courtes, la variante logarithmique rendant cette normalisation moins forte.

Alors, l'équation Eq.6.4 devient :

$$New\_overlap = \frac{overlap}{\log_2 |D(S_j)|} \quad \text{(Eq.6.5)}$$

Finalement, nous avons intégré ce module de désambiguïsation sémantique des mots arabe dans le system de recherche OIRDA pour l'ajouter l'aspect sémantique.

6.3. Conclusion

Nous avons présenté dans ce chapitre notre approche de recherche d'information sémantique sur texte arabe basée sur des connaissances, que ce soit les mesures au niveau local ou les algorithmes au niveau global. D'un point de vue local, nous avons présenté notre module de mesure de similarité sémantique basée sur DiLAC, qui est une ressource lexicale arabe développée pour cette fin.

Les expériences que nous avons effectuées ont montré que DiLAC est une ressource lexicale arabe riche fournit des définitions et des exemples d'usages et d'autre information syntaxiques et lexicales utilisé par les mesures basées basée sur le recouvrement de traits entre deux mots, et permettant de calculer le score de similarité entre des paires de mots arabes. En plus, la mesure de similarité Lesk-ar, qui est une mesure à base de traits, donne des très bons résultats, en exploitant DiLAC, cependant, elle ne s'applique pas à AWN.

Du point de vue global, et dans l'objectif d'affecter les bons sens aux mots arabes à l'échelle d'un texte, nous avons implémenté un algorithme de désambiguïsation simplifié s'appuient principalement sur la méthode originelle de Lesk [Lesk86], et adaptée aux caractéristiques constructives de DiLAC.

Conclusion générale

7. CONCLUSION GENERALE

7.1. Synthèse

Les travaux présentés dans cette thèse s'inscrivent dans le contexte de la recherche d'information sémantique dans les documents textuels en langue arabe, et plus particulièrement dans le cadre de l'indexation sémantique basée sur les ressources lexicales externes. Ces systèmes utilisent les approches de la désambiguïsation sémantique des mots pour ajouter l'aspect sémantique aux applications des dits systèmes.

Il existe différentes approches de désambiguïsation : d'une part les approches supervisées, nécessitant des corpus d'entraînement étiquetés et, d'autre part, des approches non-supervisées, qui ne nécessitent pas de corpus étiquetés, raison pour laquelle elles sont intéressantes.

Les approches non-supervisées se divisent en deux catégories : d'une part les approches non supervisées qui exploitent les données non annotées ; et d'autre part les approches à base de connaissances qui utilisent des connaissances extraites de ressources lexicales. Nous nous intéressons ici à ces dernières.

En revanche, la plupart des travaux de recherche dans le domaine de RI sémantique se sont orientés vers des documents textuels latins et peu d'études ont été effectuées sur des documents en langue arabe dont les caractéristiques grammaticales et morphologiques complexes rendent la tâche du traitement automatique encore plus difficile, plus particulièrement, pendant la lemmatisation des mots où on ne peut pas désambiguïser, dans une phase ultérieure, un lemme incorrect. En plus, la nature même des applications, qui sont souvent trop réduites ne permet pas d'envisager des solutions à grande échelle, ou encore, simplement à cause de l'indisponibilité de ressources linguistiques moderne et de corpus d'expérimentations et d'évaluations en arabe suffisantes sous forme numérique.

Nous nous sommes intéressés dans cette thèse à proposer des solutions, à répondre même à de telles problématiques, et à proposer un modèle de recherche d'information sémantique dans les documents textuels arabes.

Dans ce cadre-là, nous avons présenté principalement trois contributions traduisant nos points de vue de la lemmatisation des mots arabes, la construction d'un corpus d'expérimentation et d'évaluation en arabe, et l'utilisation d'une ressource lexicale à l'extérieur d'un système de RI dans les processus d'indexation et de recherche sémantique.

Dans notre première contribution, nous avons appliqué cinq différentes méthodes de lemmatisation pour résoudre le problème de la performance des systèmes de recherche d'information arabes, et nous avons comparé les résultats et donné une conclusion sur la méthode qui donne une meilleure performance dans la recherche d'information, permettant de mieux déterminer le lemme d'un mot, la nouvelle méthode possède une meilleure performance de recherche que les autres méthodes ; les autres méthodes ne permettent pas avec réussite de regrouper sémantiquement beaucoup de mots 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 traditionnelle, 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.

Notre deuxième contribution a pour but suggérer un nouveau type d'indexation pour, d'une part, contribuer à améliorer la qualité des systèmes de RI et à extraire et à désambiguïser les mots composés, et d'une autre part, aider à la construction des corpus d'expérimentations et d'évaluations en arabe. Nous avons proposé une méthode d'indexation qui appartient à la catégorie d'indexation semi-automatique et qui consiste en deux types d'indexations. Le premier type conduit une indexation en ligne où un document est l'unité d'indexation, et ce type d'indexation se réfère au processus d'indexation qui commence directement après l'écriture de chaque unité, ce qui permet d'aider l'expert (auteur humain du texte) à sélectionner les descripteurs appropriés pour améliorer les résultats de la recherche ; la sortie de ce processus donne lieu à un indice partiel. Le second type - selon cette méthode - est une indexation hors ligne, qui se réfère au processus d'indexation basé sur la collecte de documents textuels disponibles à partir de différents corpus, et la sortie de ce processus conduit à un index général.

Ce modèle a montré les effets de l'indexation en ligne, qui nécessite l'indexation semi-automatique, sur la performance du système de recherche d'information. En outre, ce modèle s'est avéré efficace pour aider à minimiser la taille de stockage d'index, et ainsi, améliorer le temps de réponse des différentes requêtes. Par conséquent, nous recommandons d'intégrer ce modèle dans les outils de traitement de texte afin de permettre à l'éditeur de contribuer

efficacement à la construction d'index de haute qualité tout en tenant compte des inconvénients et des faiblesses de ce modèle. Nous avons proposé également une solution aux problèmes et aux insuffisances dont souffre le traitement en langue arabe, notamment en ce qui concerne la construction de corpus en développant un cadre d'application pour la construction et le développement de corpus. En outre, nous avons suggéré une solution pour réduire les déficiences dont souffrent les systèmes d'évaluation d'extraction d'information, qui permettent aux chercheurs de tester leurs algorithmes d'indexation et de récupération, et de compléter les systèmes sur des tâches et des ensembles de données communs.

Notre troisième contribution dans cette thèse est la construction de DiLAC et l'étude de la possibilité d'appliquer une mesure de similarité sémantiques traditionnelle sur DiLAC, et cette mesure a été appliquée à l'aide d'un jeu de données de référence arabe appelé AWSS.

Les expériences que nous avons effectuées ont montré que DiLAC est une ressource lexicale arabe riche fournissant des définitions et des exemples d'usages et d'autres informations syntaxiques et lexicales utilisées par les mesures basées sur le recouvrement de traits entre deux mots, et permettant de calculer le score de similarité entre des paires de mots arabes. En plus, la mesure de similarité Lesk-ar, qui est une mesure à base de traits, donne des très bons résultats, en exploitant DiLAC, cependant, elle ne s'applique pas à AWN.

D'un point de vue global, et dans l'objectif d'affecter les bons sens aux mots arabes à l'échelle d'un texte, nous avons implémenté un algorithme de désambiguïsation simplifié qui s'appuient principalement sur la méthode originelle de Lesk [Lesk86], et s'adapte aux caractéristiques constructives de DiLAC.

7.2. Perspectives

Les perspectives envisageables à nos travaux portent principalement sur trois volets.

Un premier volet porte sur une étude plus approfondie de la construction d'un noyau sémantique pour représenter les documents en utilisant DiLAC, ainsi que l'étude de la manière d'extraire des relations sémantiques de cette ressource lexicale comme (synonymie, hyponymie, ...).

Le deuxième volet consiste à tester le modèle de RI sémantique proposé sur plusieurs collections volumineuses, afin de valider nos propositions à grande échelle et de fixer la valeur du paramètre de la formule de pondération sémantique, qui permet de balancer entre l'importance sémantique du mot dans le document et sa fréquence relative.

Enfin, le troisième volet concerne l'élargissement du modèle pour supporter une plateforme multi-sources. En intégrant plusieurs ressources sémantiques couvrant des domaines spécifiques différents, le système peut être étendu de sorte à ce qu'il détecte automatiquement la ressource adéquate pour une collection de documents donnée et propose ainsi de l'indexer via la ressource sémantique appropriée.

Références bibliographiques

REFERENCES BIBLIOGRAPHIQUES

[AAAW13] Al-Kabi, Mohammed ; Al-Belaili, Hassan ; Abul-Huda, Bilal ; Wahbeh, A: Keyword extraction based on word co-occurrence statistical information for arabic text. In: Abhath Al-Yarmouk:" Basic Science & Engineering Bd. 22 (2013), Nr. 1, S. 75–95

[AbBR08a] Abouenour, Lahsen ; Bouzoubaa, Karim ; Rosso, Paolo: Construction de l'ontologie Amine Arabic WordNet dans le cadre des systèmes Q/R. In: Proc. 2nd JOurnées Scientifiques en Technologies de l'Information et de la Communication JOSTIC-2008, Rabat, Marroco (2008)

[AbBR08b] Abouenour, Lahsen ; Bouzoubaa, Karim ; Rosso, Paolo: Improving Q/A using Arabic wordnet. In: Proc. The 2008 International Arab Conference on Information Technology (ACIT'2008), Tunisia, December, 2008

[AbBR09] Abouenour, Lahsen ; Bouzoubaa, Karim ; Rosso, Paolo: Structure-based evaluation of an Arabic semantic Query Expansion using the JIRS Passage Retrieval system. In: Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages : Association for Computational Linguistics, 2009, S. 62–68

[Abne02] Abney, Steven: Bootstrapping. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics : Association for Computational Linguistics, 2002, S. 360–367

[AbVi97] Abiteboul, Serge ; Vianu, Victor: Queries and Compuation on the Web. In: International Conference on Database Theory : Springer, 1997, S. 262–275

[ADAC16a] Abderrahim, Mohammed Alaeddine ; Dib, Mohammed ; Abderrahim, Mohammed El Amine ; Chikh, Mohammed Amine: Semantic indexing of Arabic texts for information retrieval system. In: International Journal of Speech Technology Bd. 19 (2016), Nr. 2

[ADAC16b] Abderrahim, Mohammed Alaeddine ; Dib, Mohammed ; Abderrahim, Mohammed El Amine ; Chikh, Mohammed Amine: Semantic indexing of Arabic texts for information retrieval system. In: International Journal of Speech Technology Bd. 19, Springer US (2016), Nr. 2, S. 229–236

[AEAC13] Abderrahim, Mohammed Alaeddine ; El, Mohammed ; Abderrahim, Amine ; Chikh, Mohammed Amine: Using Arabic Wordnet for semantic indexation in information retrieval system Bd. 10 (2013), Nr. 1, S. 327–332

[AgEd07] Agirre, Eneko ; Edmonds, Philip: Word sense disambiguation: Algorithms and applications. Bd. 33 : Springer Science & Business Media, 2007 — ISBN 1402048092

[AgMa01] Agirre, Eneko ; Martinez, David: Learning class-to-class selectional preferences. In: Proceedings of the 2001 workshop on Computational Natural Language Learning-Volume 7 : Association for Computational Linguistics, 2001, S. 3

[AgRi96] Agirre, Eneko ; Rigau, German: Word sense disambiguation using conceptual density. In: Proceedings of the 16th conference on Computational linguistics-Volume 1 : Association for Computational Linguistics, 1996, S. 16–22

[AhLo93] Ahlswede, Thomas E ; Lorand, David: The Ambiguity Questionnaire: A Study of Lexical Disambiguation by Human Informants. In: Proceedings of the Fifth Midwest Artificial Intelligence and Cognitive Society Conference : Chesterton Indiana, 1993, S. 21–25

[Ahls92] Ahlswede, Thomas E: Issues in the design of test data for lexical disambiguation by humans and machines. In: Proceedings of the Fourth Midwest Artificial Intelligence and Cognitive Science Society Conference, 1992, S. 112–116

[Ahls93] Ahlswede, Thomas E: Sense Disambiguation Strategies for Humans and Machines. In: Proceedings of the 9th Annual Conference on the New Oxford English Dictionary : Oxford England, 1993, S. 75–88

[Ahls95] Ahlswede, Thomas E: Word sense disambiguation by human informants. In: Proceedings of the Sixth Midwest Artificial Intelligence and Cognitive Society Conference : Carbondale Illinois, 1995, S. 73–78

[Ahme00] Ahmed, Mohamed Attia: A large-scale computational processor of the Arabic morphology, and applications. In: A Master's Thesis, Faculty of Engineering, Cairo University, Cairo, Egypt (2000)

[AlAA08] Al-Harbi, S ; Almuhareb, A ; Al-Thubaity, A: Automatic Arabic text classification. In: 9es Journées internationales Analyse statistique des Données Textuelles (2008), S. 77–84

[AlAb17] Al-Anzi, Fawaz S. ; AbuZeina, Dia: Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. In: Journal of King Saud University - Computer and Information Sciences Bd. 29, King Saud University (2017), Nr. 2, S. 189–195

[AlAt04] Al-Sulaiti, Latifa ; Atwell, Eric: Designing and developing a corpus of contemporary Arabic, In: URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.216.5091&rep=rep1&type=pdf (2004)

[AlEv94] Al-Kharashi, Ibrahim A. ; Evens, Martha W.: Comparing words, stems, and roots as index terms in an Arabic Information Retrieval system. In: Journal of the American Society for Information Science Bd. 45 (1994), Nr. 8, S. 548–560

[AlEv98] Al-Shalabi, Riyad ; Evens, Martha: A computational morphology system for Arabic. In: Proceedings of the Workshop on Computational Approaches to Semitic Languages : Association for Computational Linguistics, 1998, S. 66–72

[AlKG06] Al-Shalabi, Riyad ; Kanaan, Ghassan ; Gharaibeh, Manaf H.: Arabic Text Categorization Using kNN Algorithm. In: Proceedings of The 4th International Multiconference on Computer Science and Information Technology. Bd. 4, 2006, S. 5–7

[AmFo16] Amer, Eslam ; Foad, Khaled: Akea: an Arabic keyphrase extraction algorithm. In: International Conference on Advanced Intelligent Systems and Informatics : Springer, 2016, S. 137–146

[AmWh79] Amsler, Robert Alfred ; White, John S: Development of a computational methodology for deriving natural language semantic structures via analysis of machine-readable dictionaries : Final report on NSF project MCS77-01315. Linguistics Research Center, University of Texas, 1979

[Anth54] Anthony, Edward M: An exploratory inquiry into lexical clusters. In: American Speech Bd. 29, JSTOR (1954), Nr. 3, S. 175–180

[AOBC13] Almarsoomi, Faaza A ; O'Shea, James D ; Bandar, Zuhair ; Crockett, Keeley: AWSS: An Algorithm for Measuring Arabic Word Semantic Similarity. In: Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on : IEEE, 2013 — ISBN 1479906522, S. 504–509

[ArSH93] Arnold, Doug ; Sadler, Louisa ; Humprheys, R Lee: Evaluation: an assessment. In: Machine Translation Bd. 8, Springer (1993), Nr. 1–2, S. 1–24

[Asso93] Association française de normalisation (París, Francia): Principes généraux pour l'indexation des documents : AFNOR, 1993

[Atki87] Atkins, Beryl T S: Semantic ID tags: corpus evidence for dictionary senses. In: Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary. Bd. 1736, 1987

[Audi03] Audibert, Laurent: Étude des critères de désambiguïsation sémantique automatique: résultats sur les cooccurrences. In: 10ème conférence sur le Traitement Automatique des Langues Naturelles (TALN-2003), 2003, S. pp-35

[BaPe02] Banerjee, Satanjeev ; Pedersen, Ted: An adapted Lesk algorithm for word sense disambiguation using WordNet. In: International conference on intelligent text processing and computational linguistics : Springer, 2002, S. 136–145

[Barh60] Bar-Hillel, Y: Automatic translation of languages, advances in computers, Academic Press, New York (1960)

[BaRi99] Baeza-Yates, Ricardo ; Ribeiro-Neto, Berthier: Modern information retrieval. Bd. 463 : ACM press New York, 1999

[Barw93] Barwise, Jon: Heterogeneous reasoning. In: International Conference on Conceptual Structures : Springer, 1993, S. 64–74

[BaVC06] Bakx, Gerard Escudero ; Villodre, L M ; Claramunt, G R: Machine learning techniques for word sense disambiguation. In: Unpublished doctoral dissertation, Universitat Politecnica de Catalunya (2006)

[BeAJ10] Beseiso, Majdi ; Ahmad, Abdul Rahim ; Jais, Jamilin: Semantic Arabic search tool. In: and Knowledge Engineering Conference (STAKE 2010), 2010, S. 40

[BeBr05] Berry, Michael W ; Browne, Murray: Understanding search engines: mathematical modeling and text retrieval. Bd. 17 : Siam, 2005 — ISBN 0898718163

[BeCr92] Belkin, Nj ; Croft, Wb: Information filtering and information retrieval: two sides of the same coin? In: Communications of the ACM Bd. 29 (1992), Nr. 10, S. 1–10 — ISBN 0001-0782

[Bees01] Beesley, Kenneth R: 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. Bd. 1, 2001, S. 1–8

[BEES11] Bounhas, Ibrahim ; Elayeb, Bilel ; Evrard, Fabrice ; Slimani, Yahya: Organizing Contextual Knowledge for Arabic Text Disambiguation and Terminology Extraction. In: Knowledge Organization Bd. 38 (2011), Nr. 6

[Bees98] Beesley, KR: Arabic morphological analysis on the Internet. In: Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing, 1998

[Berr73] Berry-Rogghe, Godelieve: The computation of collocations and their relevance in lexical studies. In: The computer and literary studies, Edinburgh: Edinburgh University Press (1973), S. 103–112

[BeST07] Bessou, S ; Saadi, A ; Touahria, M: Un système d'indexation et de recherche des textes en arabe (SITRA), 1er séminaire national sur le langage naturel et l'intelligence artificielle (LANIA), Université HAssiba ben Bouali, Département d'Informatique, Chlef, Algérie (2007)

[BiCP94] Bimbot, Frédéric ; Chollet, Gérard ; Paoloni, Andrea: Assessment methodology for speaker identification and verification systems-an overview of sam-a esprit project 6819-task 2500. In: Automatic Speaker Recognition, Identification and Verification, 1994

[BoGV92] Boser, Bernhard E ; Guyon, Isabelle M ; Vapnik, Vladimir N: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory : ACM, 1992 — ISBN 089791497X, S. 144–152

[BrGY83] Brown, Gillian ; Gillian, Brown ; Yule, George: Discourse analysis : Cambridge university press, 1983 — ISBN 0521284759

[BrWi94a] Bruce, Rebecca ; Wiebe, Janyce: A new approach to word sense disambiguation. In: Proceedings of the workshop on Human Language Technology : Association for Computational Linguistics, 1994 — ISBN 1558603573, S. 244–249

[BrWi94b] Bruce, Rebecca ; Wiebe, Janyce: Word-sense disambiguation using decomposable models. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics : Association for Computational Linguistics, 1994, S. 139–146

[BSZM16] Bazzi, El ; Salim, Mohamed ; Zaki, Taher ; Mammass, Driss ; Ennaji, Abdelatif: Indexation automatique des textes arabes: état de l'art. In: E-Ti: E-Review in Technologies Information (2016), Nr. 9

[BuSa01] Buitelaar, Paul ; Sacaleanu, Bogdan: Ranking and selecting synsets by domain relevance. In: Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, NAACL 2001 Workshop : In: URL: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.3835&rep=rep1&type=pdf, 2001, S. 119–124

[CDBB09] Chen, Ping ; Ding, Wei ; Bowes, Chris ; Brown, David: A fully unsupervised word sense disambiguation method using dependency knowledge. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics : Association for Computational Linguistics, 2009 — ISBN 1932432418, S. 28–36

[Chen96] Chen, Stanley F: Building probabilistic models for natural language. In: PhD Thesis, Technical Report TR-02-96, Center for Research in Computing Technology, Harvard University. (1996)

[ChGe02] Chen, Aitao ; Gey, Fredric: Building an Arabic stemmer for information retrieval. In: TREC. Bd. 2002, 2002, S. 631–639

[ChLu85] Choueka, Yaacov ; Lusignan, Serge: Disambiguation by short contexts. In: Computers and the

Humanities Bd. 19, Springer (1985), Nr. 3, S. 147–157

[Chod98] Chodorow, M: Combining local context and wordnet sense similarity for word sense disambiguation. In: WordNet, An Electronic Lexical Database (1998)

[CHSA95] Carey, Michael J ; Haas, Laura M ; Schwarz, Peter M ; Arya, Manish ; Cody, William F ; Fagin, Ronald ; Flickner, Myron ; Luniewski, Allen W ; u. a.: Towards heterogeneous multimedia information systems: The Garlic approach. In: Research Issues in Data Engineering, 1995: Distributed Object Management, Proceedings. RIDE-DOM'95. Fifth International Workshop on : IEEE, 1995 — ISBN 0818670568, S. 124–131

[ChTa96] Chanod, Jean-Pierre ; Tapanainen, Pasi: A non-deterministic tokeniser for finite-state parsing. In: Proceedings of the Workshop on Extended finite state models of language (ECAI'96), 1996

[ClMK66] Cleverdon, C ; Mills, J ; Keen, M: Factors Determining the Performance of Indexing Systems Volume 1. Design. In: ASLIB Cranfield project Cranfield Bd. Vol 2 (1966), S. 37–59

[Coll04] Collins, Michael: Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In: New developments in parsing technology : Springer, 2004, S. 19–55

[Coop88] Cooper, William S: Getting beyond boole. In: Information Processing & Management Bd. 24, Pergamon (1988), Nr. 3, S. 243–248

[CoSm83] Cottrell, Garrison W ; Small, Steven L: A connectionist scheme for modelling word sense disambiguation. In: Cognition & Brain Theory, Lawrence Erlbaum (1983)

[CrEL03] Crestan, Éric ; El-Bèze, Marc ; de Loupy, Claude: Peut-on trouver la taille de contexte optimale en désambiguïsation sémantique. In: 10e conférence TALN, 2003, S. 85–94

[CSBA13] Cavalli-Sforza, Violetta ; Saddiki, Hind ; Bouzoubaa, Karim ; Abouenour, Lahsen ; Maamouri, Mohamed ; Goshey, Emily: Bootstrapping a wordnet for an arabic dialect from other wordnets and dictionary resources. In: Computer systems and applications (aiccsa), 2013 acs international conference on : IEEE, 2013 — ISBN 1479907928, S. 1–8

[DaBZ99] Daelemans, Walter ; Van Den Bosch, Antal ; Zavrel, Jakub: Forgetting exceptions is harmful in language learning. In: Machine learning Bd. 34, Springer (1999), Nr. 1–3, S. 11–41

[Dahl88] Dahlgren, Kathleen: Naive semantics for natural language understanding : Springer, 1988 — ISBN 0898382874

[DaOa03] Darwish, Kareem ; Oard, Douglas W: CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval : MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES, 2003

[DDFL90] Deerwester, Scott ; Dumais, Susan T ; Furnas, George W ; Landauer, Thomas K ; Harshman, Richard: Indexing by latent semantic analysis. In: Journal of the American society for information science Bd. 41, Wiley Online Library (1990), Nr. 6, S. 391–407

[DeLR77] Dempster, Arthur P ; Laird, Nan M ; Rubin, Donald B: Maximum likelihood from incomplete data via the EM algorithm. In: Journal of the royal statistical society. Series B (methodological), JSTOR (1977), S. 1–38

[DiBe12] Dilekh, Tahar ; Behloul, Ali: Implementation of a New Hybrid Method for Stemming of Arabic Text. In: International Journal of Computer Applications Bd. 46 (2012), Nr. 8, S. 14–19

[DiHJ04] Diab, Mona ; Hacioglu, Kadri ; Jurafsky, Daniel: Automatic tagging of Arabic text: From raw text to base phrase chunks. In: Proceedings of HLT-NAACL 2004: Short papers : Association for Computational Linguistics, 2004 — ISBN 1932432248, S. 149–152

[Dile11] DILEKH, Tahar: Implémentation d'un outil d'indexation et de recherche des textes en arabe, thèse de magistère, Université de Batna 2 (2011)

[Dill83] Dillon, Martin: Introduction to modern information retrieval: G. Salton and M. McGill. McGraw-Hill, New York (1983). xv+ 448 pp., $32.95 ISBN 0-07-054484-0, Pergamon (1983) — ISBN 0306-4573

[DuHe16] Duwairi, Rehab ; Hedaya, Mona: Automatic keyphrase extraction for Arabic news documents based on KEA system. In: Journal of Intelligent and Fuzzy Systems Bd. 30, IOS Press (2016), Nr. 4, S. 2101–2110

[Earl73] Earl, Lois L: Use of word government in resolving syntactic and semantic ambiguities. In: Information Storage and Retrieval Bd. 9, Elsevier (1973), Nr. 12, S. 639–664

[EBVF06] Elkateb, Sabry ; Black, William ; Vossen, Piek ; Farwell, David ; Rodríguez, H ; Pease, A ; Alkhalifa, M: Arabic WordNet and the challenges of Arabic. In: Proceedings of Arabic NLP/MT Conference, London, UK : In: URL: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.492&rep=rep1&type=pdf, 2006

[EdCo01] Edmonds, Philip ; Cotton, Scott: SENSEVAL-2: overview. In: The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems : Association for Computational Linguistics, 2001, S. 1–5

[Eijk94] Van Der Eijk, Pim: Comparative discourse analysis of parallel texts. In: arXiv preprint https://arxiv.org/abs/cmp-lg/9407022 (1994)

[ElAA15] Elabd, Emad ; Alshari, Eissa ; Abdulkader, Hatem: Semantic Boolean Arabic Information Retrieval. In: arXiv preprint https://arxiv.org/abs/1512.03167 (2015)

[ElAl12] El-Shishtawy, Tarek ; Al-sammak, Abdulwahab: Arabic Keyphrase Extraction using Linguistic knowledge and Machine Learning Techniques. In: ReCALL (2012), S. 1–8

[Elha15] El-Halees, Alaa M: Arabic text classification using maximum entropy. In: IUG Journal of Natural Studies Bd. 15 (2015), Nr. 1

[ElIs06] El-Khoribi, R ; Ismael, M: An intelligent system based on statistical learning for searching in arabic text. In: ICGST International Journal on Artificial Intelligence and Machine Learning, AIML Bd. 6, In: URL: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.4743&rep=rep1&type=pdf (2006), S. 41–47

[ElRa09] El-Beltagy, Samhaa R. ; Rafea, Ahmed: KP-Miner: A keyphrase extraction system for English and Arabic documents. In: Information Systems Bd. 34, Elsevier (2009), Nr. 1, S. 132–144 — ISBN 978-1-932432-87-9

[EnPo98] Entwisle, Jim ; Powers, David M W: The present use of statistics in the evaluation of NLP parsers. In: Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural

Language Learning : Association for Computational Linguistics, 1998 — ISBN 0725806346, S. 215–224

[EsMR00] Escudero, Gerard ; Màrquez, Lluís ; Rigau, German: Naive Bayes and exemplar-based approaches to word sense disambiguation revisited. In: arXiv preprint https://arxiv.org/abs/cs/0007011 (2000)

[EsMR04] Escudero, Gerard ; Màrquez, Lluís ; Rigau, German: TALP system for the english lexical sample task. In: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004

[FaGu06] Falquet, Gilles ; Guyot, Jacques: Construire un moteur d'indexation, Ingénierie des Systèmes d'Information, 2006, vol. 11, no 4, p. 99-131.

[Finc93] Finch, Steven: Finding structure in language, Doctoral dissertation, University of Edinburgh (1993)

[Firt75] Firth, John Rupert: Modes of meaning. Papers in Linguistics 1934-51, pages 190-215, Oxford University Press, Oxford, UK.

[FJZK12] Faaza, A ; James, D ; Zuhair, A ; Keeley, A: Arabic Word Semantic Similarity. In: World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering Bd. 6 (2012), Nr. 10, S. 2497–2505

[FoKo88] Fox, Edward A ; Koll, Matthew B: Practical enhanced Boolean retrieval: Experiences with the SMART and SIRE systems. In: Information Processing & Management Bd. 24, Elsevier (1988), Nr. 3, S. 257–267

[FrBa92] Frakes, William Bruce ; Baeza-Yates, Ricardo: Information retrieval: Data structures & algorithms. Bd. 331 : prentice Hall Englewood Cliffs, NJ, 1992

[Fuch96] Fuchs, Catherine: Les ambiguïtés du français : Ophrys, 1996 — ISBN 2708007726

[GaCY92a] Gale, William A ; Church, Kenneth W ; Yarowsky, David: Using bilingual materials to develop word sense disambiguation methods. In: Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation : In: URL: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.591.1907&rep=rep1&type=pdf, 1992, S. 101–112

[GaCY92b] Gale, William A ; Church, Kenneth W ; Yarowsky, David: One sense per discourse. In: Proceedings of the workshop on Speech and Natural Language : Association for Computational Linguistics, 1992 — ISBN 1558602720, S. 233–237

[GaCY92c] Gale, William A ; Church, Kenneth W ; Yarowsky, David: A method for disambiguating word senses in a large corpus. In: Computers and the Humanities Bd. 26, Springer (1992), Nr. 5–6, S. 415–439

[GaCY92d] Gale, William ; Church, Kenneth Ward ; Yarowsky, David: Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In: Proceedings of the 30th annual meeting on Association for Computational Linguistics : Association for Computational Linguistics, 1992, S. 249–256

[GhHF09] Gharib, Tarek Fouad ; Habib, Mena Badieh ; Fayed, Zaki Taha: Arabic Text Classification Using Support Vector Machines. In: International Journal of Computers and Their Applications Bd. 16 (2009), Nr. 4, S. 192–199

[GVCC98] Gonzalo, Julio ; Verdejo, Felisa ; Chugur, Irina ; Cigarran, Juan: Indexing with WordNet synsets can

improve text retrieval. In: arXiv preprint https://arxiv.org/abs/cmp-lg/9808002 (1998)

[HaEA11] Harrag, Fouzi ; El-Qawasmah, Eyas ; Al-Salman, Abdul Malik S.: Stemming as a feature reduction technique for Arabic text categorization. In: Proceedings of the 10th International Symposium on Programming and Systems, ISPS' 2011, 2011 — ISBN 9781457709067, S. 128–133

[HaHa14] Halliday, Michael Alexander Kirkwood ; Hasan, Ruqaiya: Cohesion in english : Routledge, 2014 — ISBN 1317869605

[Hall61] Halliday, Michael Alexander Kirkwood: Categories of the theory of grammar. In: Word Bd. 17, Taylor & Francis (1961), Nr. 2, S. 241–292

[HaOL16] Hadni, Meryeme ; Ouatik, Saïd El Alaoui ; Lachkar, Abdelmonaime: Word sense disambiguation for arabic text categorization. In: Int. Arab J. Inf. Technol. Bd. 13 (2016), Nr. 1A, S. 215–222

[HaRa05] Habash, Nizar ; Rambow, Owen: Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics : Association for Computational Linguistics, 2005, S. 573–580

[Haye77a] Hayes, Philip J: On semantic nets, frames and association, In: URL: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.75.7429&rep=rep1&type=pdf (1977)

[Haye77b] Hayes, Philip: Some association-based techniques for lexical disambiguation by machine, Doctoral dissertation, Departement de Mathematiques, Ecole Polytechnique Fédérale de Lausanne (1977)

[Hear91] Hearst, Marti: Noun homograph disambiguation using local context in large text corpora. In The Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora, Oxford. (1991), S. 185–188

[Hear94] Hearst, Marti A: Multi-paragraph segmentation of expository text. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics : Association for Computational Linguistics, 1994, S. 9–16

[Hirs92] Hirst, Graeme: Semantic interpretation and the resolution of ambiguity : Cambridge University Press, 1992 — ISBN 052142898X

[HiSt98] Hirst, Graeme ; St-Onge, David: Lexical chains as representations of context for the detection and correction of malapropisms. In: WordNet: An electronic lexical database Bd. 305 (1998), S. 305–332

[HiTh97] Hirschman, Lynette ; Thompson, Henry S: Overview of evaluation in speech and natural language processing, In: URL: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.5877&rep=rep1&type=pdf (1997)

[IdVe98] Ide, Nancy ; Veronis, Jean: Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. In: Computational Linguistics Bd. 24 (1998), Nr. 1, S. 1–40

[IdVé98] Ide, Nancy ; Véronis, Jean: Introduction to the special issue on word sense disambiguation: the state of the art. In: Computational linguistics Bd. 24, MIT Press (1998), Nr. 1, S. 2–40

[INHK13] Imam, Ibrahim ; Nounou, Nihal ; Hamouda, Alaa ; Khalek, Hebat Allah Abdul: An ontology-based summarization system for arabic documents (ossad). In: Int. J. Comput. Appl Bd. 74 (2013), Nr. 17, S. 38–43

[JäKe02] Järvelin, Kalervo ; Kekäläinen, Jaana: Cumulated gain-based evaluation of IR techniques. In: ACM Transactions on Information Systems (TOIS) Bd. 20, ACM (2002), Nr. 4, S. 422–446

[JaMF99] Jain, Anil K ; Murty, M Narasimha ; Flynn, Patrick J: Data clustering: a review. In: ACM computing surveys (CSUR) Bd. 31, Acm (1999), Nr. 3, S. 264–323

[JiCo97] Jiang, Jay J ; Conrath, David W: Semantic similarity based on corpus statistics and lexical taxonomy. In: arXiv preprint https://arxiv.org/abs/cmp-lg/9709008 (1997)

[Joac98] Joachims, Thorsten: Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning : Springer, 1998, S. 137–142

[Jone64] Jones, Karen Sparck: Synonymy and semantic classification, Ph.D. thesis, University of Cambridge, Cambridge, UK, 1964

[Jone71] Jones, K. S.: Automatic Keyword Classification for Information Retrieval, Conn.] Archon Books (1971), S. 253 — ISBN 9780208012012

[Jone97] Jones, Karen Sparck: Readings in information retrieval : Morgan Kaufmann, 1997 — ISBN 1558604545

[Jorg90] Jorgensen, Julia C: The psychological reality of word senses. In: Journal of psycholinguistic research Bd. 19, Springer (1990), Nr. 3, S. 167–190

[JoWR00] Jones, K Sparck ; Walker, Steve ; Robertson, Stephen E: A probabilistic model of information retrieval: development and comparative experiments: Part 2. In: Information processing & management Bd. 36, Elsevier (2000), Nr. 6, S. 809–840

[JuMa14] Jurafsky, Dan ; Martin, James H: Speech and language processing. Bd. 3 : Pearson London, 2014

[KaAu13] Kaye, K ; Aung, Win Thandar: Word Sense Disambiguation: A Briefly Survey. In: IJCCER Bd. 1 (2013), Nr. 4, S. 118–123

[Kabb06] Kabbaj, Adil: Development of intelligent systems and multi-agents systems with amine platform. In: International Conference on Conceptual Structures : Springer, 2006, S. 286–299

[Kadr08] Kadri, Youssef: Recherche d'information translinguistique sur les documents en arabe. In: A Ph.D Thesis, Departement d'informatique et de recherche operationnelle Faculte des arts et des sciences, Universite de Montreal, Montreal, CANADA (2008)

[KaNi06a] Kadri, Youssef ; Nie, J.Y.: Effective stemming for Arabic information retrieval. In: Proceedings of the challenge of arabic for nLP/mt, international conf. at the British computer Society (BcS) (2006), S. 68–74

[KaNi06b] Kadri, Youssef ; Nie, J.Y.: Effective stemming for Arabic information retrieval. In: Proceedings of the challenge of arabic for nLP/mt, international conf. at the British computer Society (BcS) (2006), S. 68–74

[Kapl55] Kaplan, Abraham: An Experiment Study of Ambiguity and Context. In: Mechanical Translation Bd. 2 (1955), S. 39–46

[KeSt75] Kelly, Edward F ; Stone, Philip J: Computer recognition of English word senses. Bd. 13 : North-Holland, 1975 — ISBN 0444108319

[KhGa99] Khoja, Shereen ; Garside, Roger: Stemming arabic text. In: Lancaster, UK, Computing Department, Lancaster University (1999)

[Khre06] Khreisat, Laila: Arabic text classification using N-gram frequency statistics a comparative study. In: Conference on Data Mining DMIN'06 Bd. 2006 (2006), S. 78–82

[KiMr85] Kintsch, Walter ; Mross, Ernest F: Context effects in word identification. In: Journal of memory and language Bd. 24, In: URL: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.20.6790&rep=rep1&type=pdf (1985), Nr. 3, S. 336–349

[KoMa00] Kowalski, Gerald J ; Maybury, Mark T: Information storage and retrieval systems: theory and implementation. Bd. 8 : Springer Science & Business Media, 2000 — ISBN 0792379241

[KrCr92] Krovetz, Robert ; Croft, W Bruce: Lexical ambiguity and information retrieval. In: ACM Transactions on Information Systems (TOIS) Bd. 10, ACM (1992), Nr. 2, S. 115–141

[Krov98] Krovetz, Robert: More than one sense per discourse. In: NEC Princeton NJ Labs., Research Memorandum Bd. 23 (1998)

[KuMa01] Kudo, Taku ; Matsumoto, Yuji: Chunking with support vector machines. In: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies : Association for Computational Linguistics, 2001, S. 1–8

[LaBC02] Larkey, Leah S ; Ballesteros, Lisa ; Connell, Margaret E: Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval : ACM, 2002 — ISBN 1581135610, S. 275–282

[LaCo01a] Larkey, Leah S ; Connell, Margaret E: Arabic Information Retrieval at UMass in TREC-10. In: Proceedings of the 10th Text Retrieval Conference (TREC'01) (2001), Nr. Lm

[LaCo01b] Larkey, Leah S ; Connell, Margaret E: Arabic Information Retrieval at UMass in TREC-10. In: TREC, 2001

[LaGi98] Lawrence, Steve ; Giles, C Lee: Searching the world wide web. In: Science Bd. 280, American Association for the Advancement of Science (1998), Nr. 5360, S. 98–100

[LaWa93] Lancaster, F Wilfrid ; Warner, Amy J: Information Retrieval Today. Revised, Retitled : ERIC, 1993 — ISBN 0878150641

[LBEE13] Lahbib, Wiem ; Bounhas, Ibrahim ; Elayeb, Bilel ; Evrard, Fabrice ; Slimani, Yahya: A Hybrid Approach for Arabic Semantic Relation Extracion. In: Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference (2013), S. 315–320 — ISBN 9781577356059

[LeCh98] Leacock, Claudia ; Chodorow, Martin: Combining local context and WordNet similarity for word sense identification. In: WordNet: An electronic lexical database Bd. 49 (1998), Nr. 2, S. 265–283

[LeMC98] Leacock, Claudia ; Miller, George A ; Chodorow, Martin: Using corpus statistics and WordNet relations for sense identification. In: Computational Linguistics Bd. 24, MIT Press (1998), Nr. 1, S. 147–165

[LeNC04] Lee, Yoong Keok ; Ng, Hwee Tou ; Chia, Tee Kiah: Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004

[Lenc08] Lenci, Alessandro: Distributional semantics in linguistic and cognitive research. In: Italian journal of linguistics Bd. 20 (2008), Nr. 1, S. 1–31

[LeNg02] Lee, Yoong Keok ; Ng, Hwee Tou: An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 : Association for Computational Linguistics, 2002, S. 41–48

[LeRO96] Levy, Alon ; Rajaraman, Anand ; Ordille, Joann: Querying heterogeneous information sources using source descriptions. In: Proceedings 22th Int. Conf. on Very Large Data Bases, 1996, pages 251–262

[LeSh04] Le, Cuong Anh ; Shimazu, Akira: High WSD accuracy using Naive Bayesian classifier with rich features. In: Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation, 2004, S. 105–114

[Lesk86] Lesk, Michael: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the 5th annual international conference on Systems documentation : ACM, 1986 — ISBN 0897912241, S. 24–26

[LeTV93a] Leacock, Claudia ; Towell, Geoffrey ; Voorhees, Ellen: Towards building contextual representations of word senses using statistical models. In: Acquisition of Lexical Knowledge from Text (1993)

[LeTV93b] Leacock, Claudia ; Towell, Geoffrey ; Voorhees, Ellen: Corpus-based statistical sense resolution. In: Proceedings of the workshop on Human Language Technology : Association for Computational Linguistics, 1993 — ISBN 1558603247, S. 260–265

[Lin98a] Lin, Dekang: Automatic retrieval and clustering of similar words. In: Proceedings of the 17th international conference on Computational linguistics-Volume 2 : Association for Computational Linguistics, 1998, S. 768–774

[Lin98b] Lin, Dekang: Extracting collocations from text corpora. In: First workshop on computational terminology : In: URL: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.11.7962&rep=rep1&type=pdf, 1998, S. 57–63

[Luhn57] Luhn, H P: 1 A Statistical Approach to Mechanized Encoding. In: IBM journal (1957)

[Luhn58] Luhn, Hans Peter: Review of information retrieval methods. In: Claire K. Schultz (Ed.), (1968). H.P. Luhn: Pioneer of information science: Selected works (pp. 140-144). New York: Spartan books, 1958

[MaCa00] Magnini, Bernardo ; Cavaglia, Gabriela: Integrating Subject Field Codes into WordNet. In: LREC, 2000, S. 1413–1418

[MaKu60] Maron, M. E. ; Kuhns, J. L.: On Relevance, Probabilistic Indexing and Information Retrieval. In: Journal of the ACM Bd. 7, ACM (1960), Nr. 3, S. 216–244 — ISBN 0004-5411

[Marc91] Marcus, Richard S: Computer and Human Understanding in Intelligent Retrieval Assistance. In:

Proceedings of the ASIS Annual Meeting. Bd. 28 : ERIC, 1991 — ISBN 0044-7870, S. 49–59

[Mast57] Masterman, Margaret: The thesaurus in syntax and semantics. In: Mechanical Translation Bd. 4 (1957), S. 1–2

[MBSC97] Mitra, Mandar ; Buckley, Chris ; Singhal, Amit ; Cardie, Claire: An analysis of statistical and syntactic phrases. In: Computer-Assisted Information Searching on Internet : LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 1997, S. 200–214

[McCa03] McCarthy, Diana ; Carroll, John: Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. In: Computational Linguistics Bd. 29, MIT Press (2003), Nr. 4, S. 639–654

[Mcro92] McRoy, Susan W: Using multiple knowledge sources for word sense discrimination. In: Computational Linguistics Bd. 18, MIT Press (1992), Nr. 1, S. 1–30

[MEMR07] Màrquez, Lluís ; Escudero, Gerard ; Martínez, David ; Rigau, German: Supervised corpus-based methods for WSD. In: Word Sense Disambiguation : Springer, 2007, S. 167–216

[Merh09] Merhben, Laroussi: Ambiguous Arabic Words Disambiguation : The results. In: Contexts, 2009, S. 45–52

[MeZZ09] Merhben, Laroussi ; Zouaghi, Anis ; Zrigui, Mounir: Ambiguous Arabic Words Disambiguation: The results. In: Proceedings of the Student Research Workshop, 2009, S. 45–52

[MHDH08] Mansour, Nashat ; Haraty, Ramzi A. ; Daher, Walid ; Houri, Manal: An auto-indexing method for Arabic text. In: Information Processing and Management Bd. 44, Elsevier (2008), Nr. 4, S. 1538–1545

[MiCh91] Miller, George A ; Charles, Walter G: Contextual correlates of semantic similarity. In: Language and cognitive processes Bd. 6, Taylor & Francis (1991), Nr. 1, S. 1–28

[MiCS06] Mihalcea, Rada ; Corley, Courtney ; Strapparava, Carlo: Corpus-based and knowledge-based measures of text semantic similarity. In: AAAI. Bd. 6, 2006, S. 775–780

[Mill95] Miller, George A: WordNet: a lexical database for English. In: Communications of the ACM Bd. 38, ACM (1995), Nr. 11, S. 39–41

[MiMo00] Mihalcea, Rada ; Moldovan, Dan: Semantic indexing using WordNet senses. In: Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 11 : Association for Computational Linguistics, 2000, S. 35–45

[MiTF04] Mihalcea, Rada ; Tarau, Paul ; Figa, Elizabeth: PageRank on semantic networks, with application to word sense disambiguation. In: Proceedings of the 20th international conference on Computational Linguistics : Association for Computational Linguistics, 2004, S. 1126

[MKWC04] McCarthy, Diana ; Koeling, Rob ; Weeds, Julie ; Carroll, John: Finding predominant word senses in untagged text. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics : Association for Computational Linguistics, 2004, S. 279

[MoAA10] Moawad, Ibrahim Fathy ; Abdeen, Mohammad ; Aref, Mostafa Mahmoud: Ontology-based

architecture for an arabic semantic search engine. In: The Tenth Conference. On Language Engineering Organized by Egyptian Society of Language Engineering (ESOLEC'2010), 2010, S. 15–16

[MoHA12] Molijy, Abdulrahman Al ; Hmeidi, Ismail ; Alsmadi, Izzat: Indexing of Arabic documents automatically based on lexical analysis. In: International Journal on Natural Language Computing (IJNLC) Bd. 1 (2012), Nr. 1, S. 1–8

[MoHi91] Morris, Jane ; Hirst, Graeme: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. In: Computational linguistics Bd. 17, MIT Press (1991), Nr. 1, S. 21–48

[MoMo02] Mohamadi, T ; Mokhnache, S: Design and development of Arabic speech synthesis, WSEAS 2002, Greece, 2002

[MoMo17] Mohammed, Nababteh ; Mohammed, Deri: Experimental Study of Semantic Similarity Measures on Arabic WordNet. In: International Journal of Computer Science and Network Security (IJCSNS) Bd. 17, International Journal of Computer Science and Network Security (2017), Nr. 2, S. 131

[Mooe50] Mooers, Calvin S: EDITOR'S CORNER:" Coding, Information Retrieval, and the Rapid Selector". In: Journal of the American Society for Information Science Bd. 1, American Documentation Institute (1950), Nr. 4, S. 225

[Morr88] Morris, Jane: Lexical cohesion, the thesaurus, and the structure of text, Computer Systems Research Institute, University of Toronto (1988)

[MoWa10] Mohamed, R. ; Watada, J.: An evidential reasoning based LSA approach to document classification for knowledge acquisition. In: IEEM2010 - IEEE International Conference on Industrial Engineering and Engineering Management, 2010 — ISBN 9781424485031, S. 1092–1096

[MRRZ14] Mahgoub, Ashraf ; Rashwan, Mohsen ; Raafat, Hazem ; Zahran, Mohamed ; Fayek, Magda: Semantic query expansion for Arabic information retrieval. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 2014, S. 87–92

[MSPG01] Magnini, Bernardo ; Strapparava, Carlo ; Pezzulo, Giovanni ; Gliozzo, Alfio: Using domain information for word sense disambiguation. In: The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems : Association for Computational Linguistics, 2001, S. 111–114

[MUUM01] Murata, Masaki ; Utiyama, Masao ; Uchimoto, Kiyotaka ; Ma, Qing ; Isahara, Hitoshi: Japanese word sense disambiguation using the simple Bayes and support vector machine methods. In: The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems : Association for Computational Linguistics, 2001, S. 135–138

[Navi09] Navigli, Roberto: Word sense disambiguation: A survey. In: ACM computing surveys (CSUR) Bd. 41, ACM (2009), Nr. 2, S. 10

[NeSh05] Nelken, Rani ; Shieber, Stuart M: Arabic diacritization using weighted finite-state transducers. In: Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages : Association for Computational Linguistics, 2005, S. 79–86

[Ng97] Ng, Hwee Tou: Exemplar-based word sense disambiguation: Some recent improvements. In: arXiv

preprint https://arxiv.org/abs/cmp-lg/9706010 (1997)

[Oswa52] Oswald Jr, Victor A: Microsemantics. In: Communication presented at the first MIT conference on Mechanical Translation, 1952, S. 17–20

[PaBP03] Patwardhan, Siddharth ; Banerjee, Satanjeev ; Pedersen, Ted: Using measures of semantic relatedness for word sense disambiguation. In: International Conference on Intelligent Text Processing and Computational Linguistics : Springer, 2003, S. 241–257

[Patw03] Patwardhan, Siddharth: Incorporating dictionary and corpus information into a context vector measure of semantic relatedness, Master's thesis, University of Minnesota, Duluth (2003)

[PeBr97] Pedersen, Ted ; Bruce, Rebecca: Distinguishing word senses in untagged text. In: arXiv preprint https://arxiv.org/abs/cmp-lg/9706008 (1997)

[PeBr98] Pedersen, Ted ; Bruce, Rebecca: Knowledge lean word-sense disambiguation. In: AAAI/IAAI, 1998, S. 800–805

[Pede00] Pedersen, Ted: A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference : Association for Computational Linguistics, 2000, S. 63–69

[Pede01] Pedersen, Ted: A decision tree of bigrams is an accurate predictor of word sense. In: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies : Association for Computational Linguistics, 2001, S. 1–8

[Powe98] Powers, David M W: Applications and explanations of Zipf's law. In: Proceedings of the joint conferences on new methods in language processing and computational natural language learning : Association for Computational Linguistics, 1998 — ISBN 0725806346, S. 151–160

[Pust91] Pustejovsky, James: The generative lexicon. In: Computational linguistics Bd. 17, MIT press (1991), Nr. 4, S. 409–441

[QRSU95] Quass, Dallan ; Rajaraman, Anand ; Sagiv, Yehoshua ; Ullman, Jeffrey ; Widom, Jennifer: Querying semistructured heterogeneous information. In: International Conference on Deductive and Object-Oriented Databases : Springer, 1995, S. 319–344

[Quin14] Quinlan, J Ross: C4. 5: programs for machine learning : Elsevier, 2014 — ISBN 0080500587

[RaDi10] Raheel, Saeed ; Dichy, Joseph: An empirical study on the feature's type effect on the automatic classification of Arabic documents. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Bd. 6008 LNCS : Springer, 2010 — ISBN 3642121152, S. 673–686

[RaMo91] Rayner, Keith ; Morris, Robin K: Comprehension processes in reading ambiguous sentences: Reflections from eye movements. In: Advances in psychology. Bd. 77 : Elsevier, 1991 — ISBN 0166-4115, S. 175–198

[Reif54] Reifler, Erwin: The first conference on mechanical translation. In: Mechanical Translation Bd. 1 (1954),

Nr. 2, S. 23–32

[Resn95] Resnik, Philip: Using information content to evaluate semantic similarity in a taxonomy. In: arXiv preprint https://arxiv.org/abs/cmp-lg/9511007 (1995)

[Resn97a] Resnik, Philip: Selectional preference and sense disambiguation. In: Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics, Washington, D.C. 1997

[Resn97b] Resni, Philip: A perspective on word sense disambiguation methods and their evaluation. In: ACL-SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What, and How? Washington, D.C., pp 79 - 86. 1997

[Rijs79] Van Rijsbergen, C J: Information retrieval. dept. of computer science, university of glasgow. In: URL: citeseer.ist. psu. edu/vanrijsbergen79information. html Bd. 14 (1979)

[RiSA97] Richmond, Korin ; Smith, Andrew James ; Amitay, Einat: Detecting subject boundaries within text: A language-independent statistical approach. In: Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, EMNLP-2, pages 47-54, Brown University, Providence, RI, 1997

[RMBB89] Rada, Roy ; Mili, Hafedh ; Bicknell, Ellen ; Blettner, Maria: Development and application of a metric on semantic nets. In: IEEE transactions on systems, man, and cybernetics Bd. 19, IEEE (1989), Nr. 1, S. 17–30

[RoJo76] Robertson, Stephen E ; Jones, K Sparck: Relevance weighting of search terms. In: Journal of the American Society for Information science Bd. 27, Wiley Online Library (1976), Nr. 3, S. 129–146

[RuGo65] Rubenstein, Herbert ; Goodenough, John B: Contextual correlates of synonymy. In: Communications of the ACM Bd. 8, ACM (1965), Nr. 10, S. 627–633

[SaBu90] Salton, Gerard ; Buckley, Chris: Improving retrieval performance by relevance feedback. In: Journal of the American society for information science Bd. 41, Wiley Online Library (1990), Nr. 4, S. 288–297

[SaFW83] Salton, Gerard ; Fox, Edward A ; Wu, Harry: Extended Boolean information retrieval. In: Communications of the ACM Bd. 26, ACM (1983), Nr. 11, S. 1022–1036

[Salt68] Salton, Gerard: Automatic information organization and retrieval, McGraw-Hill Book Company, New York, 1968

[SaMc86] Salton, Gerard ; McGill, Michael J: Introduction to modern information retrieval, McGraw-Hill Book Company, New York, 1986

[Sand94] Sanderson, Mark: Word sense disambiguation and information retrieval. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval : Springer-Verlag New York, Inc., 1994 — ISBN 038719889X, S. 142–151

[SaWY75] Salton, Gerard ; Wong, Anita ; Yang, Chung-Shu S.: A vector space model for automatic indexing. In: Communications of the ACM Bd. 18, ACM (1975), Nr. 11, S. 613–620 — ISBN 0001-0782

[Schü98] Schütze, Hinrich: Automatic word sense discrimination. In: Computational linguistics Bd. 24, MIT Press (1998), Nr. 1, S. 97–123

[ScPe95] Schütze, Hinrich ; Pedersen, Jan O: Information retrieval based on word senses, In: URL:

citeseer.ist.psu.edu/viewdoc/download?doi=10.1.1.11.8934&rep=rep1&type=pdf (1995)

[SGNB14] Singh, Richard Laishram ; Ghosh, Krishnendu ; Nongmeikapam, Kishorjit ; Bandyopadhyay, Sivaji: A decision tree based word sense disambiguation system in manipuri language. In: Advanced Computing Bd. 5, Academy & Industry Research Collaboration Center (AIRCC) (2014), Nr. 4, S. 17

[ShTM15] Shutova, Ekaterina ; Tandon, Niket ; De Melo, Gerard: Perceptually grounded selectional preferences. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Bd. 1, 2015, S. 950–960

[SiMi07] Sinha, Ravi ; Mihalcea, Rada: Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In: Semantic Computing, 2007. ICSC 2007. International Conference on : IEEE, 2007 — ISBN 0769529976, S. 363–369

[Skor72] SKOROKHOD, K O: Adaptive method of automatic abstracting and indexing. In: Information Processing 71, North Holland (1972), S. 1179–1182

[SLPW99] Strzalkowski, Tomek ; Lin, Fang ; Perez-Carballo, Jose ; Wang, Jin: Evaluating Natural Language Processing Techniques in Information Retrieval: a TREC Perspective. In: Natural Language : Springer, 1999, S. 1–26

[Spoe93] Spoerri, Anselm: InfoCrystal: A visual tool for information retrieval & management. In: Proceedings of the second international conference on Information and knowledge management : ACM, 1993 — ISBN 0897916263, S. 11–20

[StPo98] Steele, Robert ; Powers, David: Evolution and evaluation of document retrieval queries. In: Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning : Association for Computational Linguistics, 1998 — ISBN 0725806346, S. 163–164

[Strz99] Strzalkowski, Tomek: Natural Language Information Retrieval. Bd. 7 : Springer Science & Business Media, 1999 — ISBN 0792356853

[StWi01] Stevenson, Mark ; Wilks, Yorick: The interaction of knowledge sources in word sense disambiguation. In: Computational Linguistics Bd. 27, MIT Press (2001), Nr. 3, S. 321–349

[TaAl12] Tahar, Dilekh ; Ali, Behloul: Implementation of a New Hybrid Method for Stemming of Arabic Text. In: 1st Taibah University International Conference on Computing and Information Technology, ICCIT 2012, Al Madinah, kingdom of Saudi Arabia. : Proceedings of ICCIT, 2012, 580–585

[Tabo89] Tabossi, Patrizia: What's in a context? In: Resolving semantic ambiguity : Springer, 1989, S. 25–39

[Tabo91] Tabossi, Patrizia: Understanding words in context. In: Advances in psychology. Bd. 77 : Elsevier, 1991 — ISBN 0166-4115, S. 1–22

[Taha76] Tahani, Valiollah: A fuzzy model of document retrieval systems. In: Information Processing & Management Bd. 12, Elsevier (1976), Nr. 3, S. 177–187

[TaSA18a] Tahar, Dilekh ; Saber, Benharzallah ; Ali, Behloul: The real-time indexing of Arabic documents and

their role in improving the accuracy of information retrieval. In: 1st Joint International Conference_ QU_WORAL_ Computational Linguistics and Arabic Language Processing, Doha, Qatar., 2018 — ISBN 1479906522

[TaSA18b] Tahar, Dilekh ; Saber, Benharzallah ; Ali, Behloul: The Impact of Online Indexing in Improving Arabic Information Retrieval Systems. In: Informatica Bd. 42 (2018), Nr. 4

[TaYB09] Tazzite, N ; Yousfi, A ; Bouyakhef, E H: Design and implementation of an information retrieval system by integrating semantic knowledge in the indexing phase. In: Artificial Intelligence and Machine Learning AIML Bd. 9 (2009), Nr. 1, S. 49–56

[TEZH09] Thabtah, Fadi ; Eljinini, Mohammad Ali H ; Zamzeer, Mannam ; Hadi, Musa: Naïve Bayesian Based on Chi Square to Categorize Arabic Data. In: Communications Bd. 10 (2009), S. 158–163

[Thab08] Thabtah, Fadi: VSMs with K-Nearest Neighbour to categorise Arabic text data. In: Proceedings of the World Congress on Engineering and Computer Science (2008), Nr. WCECS 2008, October 22-24, 2008, San Francisco, USA, S. 22–25 — ISBN 9789889867102

[THYB07] Tazit, Naima ; El Hossin Bouyakhf, Souad Sabri ; Yousfi, Abdellah ; Bouzouba, Karim: Semantic internet search engine with focus on Arabic language, In: URL: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.547.4822&rep=rep1&type=pdf (2007)

[TlMe06] Tlili-Guiassa, Yamina ; Merouani, Hayet Farida: Désambiguïsation sémantique d'un texte Arabe. In: Conference TALN. Bd. 6, 2006

[Trot05] Trotman, Andrew: Learning to rank. In: Information Retrieval Bd. 8, Springer (2005), Nr. 3, S. 359–381

[TuCr91] Turtle, Howard ; Croft, W Bruce: Evaluation of an inference network-based retrieval model. In: ACM Transactions on Information Systems (TOIS) Bd. 9, ACM (1991), Nr. 3, S. 187–222

[TYBM00] Tikk, Domonkos ; Yang, Jae Dong ; Biró, György ; Muresan, Leila: Text categorization with hierarhical category structure. In: Proc. of the 3rd Int. Conf. of Hungarian Researchers in Computational Intelligence (HUCI'02), S. 85

[Vapn98] Vapnic, Vladimir N: Statistical learning theory. In: A Wiley-Interscience Publication (1998)

[Véro03] Véronis, Jean: Cartographie lexicale pour la recherche d'information. In: Actes de TALN 2003 (2003), S. 265–274

[Véro04] Véronis, Jean: Hyperlex: lexical cartography for information retrieval. In: Computer Speech & Language Bd. 18, Elsevier (2004), Nr. 3, S. 223–252

[Voor93] Voorhees, Ellen M: Using WordNet to disambiguate word senses for text retrieval. In: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval : ACM, 1993 — ISBN 0897916050, S. 171–180

[Voor95] Voorhees, Ellen M: Learning context to disambiguate word senses. In: Selecting Good Models, MIT Press (1995)

[WaPo85] Waltz, David L ; Pollack, Jordan B: Massively parallel parsing: A strongly interactive model of natural

language interpretation. In: Cognitive science Bd. 9, Wiley Online Library (1985), Nr. 1, S. 51–74

[Weav49] WEAVER, Warren: Translation. Reprinted in Readings in Machine Translation, Nirenburg et al.(eds.)(2003), MIT Press (1949)

[WFHP16] Witten, Ian H ; Frank, Eibe ; Hall, Mark A ; Pal, Christopher J: Data Mining: Practical machine learning tools and techniques : Morgan Kaufmann, 2016 — ISBN 0128043571

[WiFr00] Witten, Ian H ; Frank, Eibe: Weka. In: Machine Learning Algorithms in Java (2000), S. 265–320

[WiGa07] Wightwick, Jane ; Gaafar, Mahmoud: Arabic Verbs and Essentials of Grammar, 2E (Verbs and Essentials of Grammar Series). Bd. 71498052 : McGraw-Hill Companies, Inc., ISBN-10, 2007

[Wilk72] Wilks, Yorick: An artificial intelligence approach to machine translation : In: Roger Schank and Kenneth Colby, editors, Computer Models of Thought and Language W. H. Freeman, San Francisco, pages 114-151, 1972

[Wilk75] Wilks, Yorick: A preferential, pattern-seeking, semantics for natural language inference. In: Artificial intelligence Bd. 6, Elsevier (1975), Nr. 1, S. 53–74

[Wilk93] Wilks, Yorik: Providing machine tractable dictionary tools. In: Semantics and the Lexicon : Springer, 1993, S. 341–401

[WiSt96] Wilks, Yorick ; Stevenson, Mark: The Grammar of Sense: Is word-sense tagging much more than part-of-speech tagging? In: arXiv preprint https://arxiv.org/abs/cmp-lg/9607028 (1996)

[WuPa94] Wu, Zhibiao ; Palmer, Martha: Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics : Association for Computational Linguistics, 1994, S. 133–138

[Yaro92] Yarowsky, David: Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In: Proceedings of the 14th conference on Computational linguistics-Volume 2 : Association for Computational Linguistics, 1992, S. 454–460

[Yaro93] Yarowsky, David: One sense per collocation. In: Proceedings of the workshop on Human Language Technology : Association for Computational Linguistics, 1993 — ISBN 1558603247, S. 266–271

[Yaro94] Yarowsky, David: Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics : Association for Computational Linguistics, 1994, S. 88–95

[Yaro95a] Yarowsky, David: Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd annual meeting on Association for Computational Linguistics : Association for Computational Linguistics, 1995, S. 189–196

[Yaro95b] Yarowsky, David Eric: Three machine learning algorithms for lexical ambiguity resolution. In: PhD Thesis, Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA, USA. (1995)

[Yaro99] Yarowsky, David: A comparison of corpus-based techniques for restoring accents in Spanish and French

text. In: Natural language processing using very large corpora : Springer, 1999, S. 99–120

[Ye04] Ye, Mr: Selectional Preferenced Based Verb Sense Disambiguation Using WordNet, In: Ash Asudeh, Cécile Paris & Stephen Wan, Proceedings of the Australasian Language Technology Workshop, 155–162, 2004

[ZENM10] ZAKI, Taher ; ENNAJI, Abdellatif ; MAMMASS, Driss: A semantic proximity based system of Arabic text indexation. In : International Conference on Image and Signal Processing. Springer, Berlin, Heidelberg, p. 419-427, 2010

[ZhGo09] Zhu, Xiaojin ; Goldberg, Andrew B: Introduction to semi-supervised learning. In: Synthesis lectures on artificial intelligence and machine learning Bd. 3, Morgan & Claypool Publishers (2009), Nr. 1, S. 1–130

[ZoMR98] Zobel, Justin ; Moffat, Alistair ; Ramamohanarao, Kotagiri: Inverted files versus signature files for text indexing. In: ACM Transactions on Database Systems (TODS) Bd. 23, ACM (1998), Nr. 4, S. 453–490

[ZoMZ12] Zouaghi, Anis ; Merhbene, Laroussi ; Zrigui, Mounir: Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation. In: Artificial Intelligence Review Bd. 38 (2012), Nr. 4, S. 257–269

[ZZAM12] Zouaghi, Anis ; Zrigui, Mounir ; Antoniadis, Georges ; Merhbene, Laroussi: Contribution to Semantic Analysis of Arabic Language. In: Advances in Artificial Intelligence Bd. 2012 (2012), S. 1–8

[مختا08] مختار, عمر، أحمد: معجم اللغة العربية المعاصرة : Bd. 4 .عالم الكتب، 2008