الجـــــــــــــــــــمـهـوريـــــة الجـــــــــــــزائـريـة الديمـــــــقـراطيـة الـشـــــــــعبيـة

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التـــــــــــــــــــــــــــعـليـم العـــــــالي و البـــــــــحث العـــــلـمـــي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جـــــــــــــــامعة أبي بكـــر بـلـقـايد – تـلمســـــــــــــان –

Université Aboubakr Belkaïd – Tlemcen –
Faculté de SCIENCES

THESE

Présentée pour l'obtention du **grade** de **DOCTEUR EN SCIENCES**

En : Informatique

Par : ABDERRAHIM Mohammed Alaeddine

Sujet

Exploitation des Ontologies dans les Systèmes de recherche d'informations Arabes

Soutenue publiquement, le 25/ 02 / 2016 , devant le jury composé de :

Mr LEHSAINI Mohammed	MCA	Université de Tlemcen	Président
Mr ABDERRAHIM Mohammed El Amine	MCA	Université de Tlemcen	Directeur de thèse
Mr CHIKH Mohammed Amine	Professeur	Université de Tlemcen	Examinateur
Mr LEHIRECHE Ahmed	Professeur	Université de Sidi Bel Abbes	Examinateur
Mme HAMDADOU Djamila	MCA	Université d'Oran	Examinatrice
Mr BOUCHIHA Djalloul	MCA	Centre Universitaire NAAMA	Examinateur

Table des matières

Chapitre 3 : La Recherche Sémantique pour les Textes Arabes : Etat de l'Art

Chapitre 4 : La Recherche Sémantique pour les Textes Arabes : Contribution

Remerciements

Je tiens tout d'abord à remercier Monsieur ABDERRAHIM Mohammed El-Amine, Maitre de conférences à l'université de Tlemcen, pour la confiance qu'il m'a accordé en acceptant de diriger cette thèse et sa patience de la suivre jusqu'à son aboutissement. Je le remercie profondément pour son attention, sa bienveillance et son appui sans faille qui ont été des encouragements décisifs pour mener à terme ce travail. Sans ses qualités rares au niveau humain et scientifique, le développement et l'achèvement de ce travail n'auraient été possibles. Je suis sincèrement reconnaissant à vous, Monsieur ABDERRAHIM. J'ai un grand honneur et une grande chance d'avoir un encadreur comme vous.

Je voudrais également remercier les personnes qui me font l'honneur d'accepter de participer au jury de cette thèse :

- Monsieur LEHSAINI Mohammed, Maitre de conférences à l'université de Tlemcen. Qu'il trouve ici toute ma reconnaissance pour avoir accepté de présider le jury de cette thèse.
- Monsieur CHIKH Mohammed Amine, Professeur à l'université de Tlemcen, pour l'intérêt qu'il a porté à mon travail et pour le temps qu'il a consacré en acceptant d'être examinateur et de participer au jury.
- J'adresse également mes remerciements à Monsieur LEHIRECHE Ahmed, Professeur à l'université de Sidi Bel Abbes qui a accepté d'être examinateur et de participer au jury.
- J'adresse également mes remerciements à Madame HAMDADOU Djamila, Maitre de conférences à l'université d'Oran qui a accepté d'être examinatrice et de participer au jury.
- J'adresse également mes remerciements à Monsieur BOUCHIHA Djalloul, Maitre de conférences au centre universitaire NAAMA qui a accepté d'être examinateur et de participer au jury.

Je tiens aussi à mentionner le plaisir que j'ai eu à travailler au sein de l'université d'Ibn Khaldoun de Tiaret et à l'université de Tlemcen, et j'en remercie ici tous les collègues, pour leur gentillesse, leur amitié.

Malgré que tous les mots restent faibles pour exprimer mes sentiments, qu'ils trouvent à travers ce travail les fruits et la récompense de leurs efforts. Je remercie mes parents pour leur soutien moral, spirituel et leur tolérance durant toutes mes années d'études. J'espère que le dieu me donne la force et le courage pour que je puisse rendre leurs sacrifices.

Mes remerciements vont aussi à mes proches et aux membres de ma famille en particulier mon frère et ma sœur, dont les encouragements et le soutien ont été indispensables à l'aboutissement de mes études.
Je voudrais remercier tous ceux qui ont facilité ma tâche et m'ont permis de mener à bien ce travail ainsi que ceux qui m'ont aidé dans mes études, et que je n'ai pas pu citer.
Enfin, je remercie Dieu tout puissant de la patience et de la volonté qu'il m'a donné pour réaliser ce travail.

Liste des Tableaux

Liste des Figures

ACRONYME

Arabic Dictionary of Meaning	ADM
British National Corpus	BNC
Concept Frequency	CF
Collins English Dictionnary	CED
Indexation Sémantique	IS
Indexation Conceptuelle	IC
Inverse of Document Frequency	IDF
Longman Dictionary of Contemporary English	LDOCE
Medical Subject Heading	MESH
Mot Graphique	MG
Mot Graphique Arabe	MGA
Système de Recherche d'Information	SRI
Recherche d'Information	RI
Term Frequency	TF
Questions/Réponses	QR
WordNet Arabe	AWN
Word Sense Disambiguation	WSD

Introduction générale

De nos jours, l'information est devenue disponible en grande quantité et en plusieurs format (image, son, vidéo, texte). Cette quantité énorme doit être accessible et contrôlable par la pluparts des utilisateurs qui veulent accéder et manipuler ces informations. Afin de permettre à une personne d'accéder à cette information, un Système de Recherche d'Information (SRI) est met en disposition. Ce système a pour but d'organiser et de faciliter la manipulation d'information. La Recherche d'Information (RI) s'intéresse principalement à sélectionner à partir d'un ensemble de documents existants, ceux qui sont pertinents à une requête utilisateur. Une définition claire est donnée par Lancaster: "la RI est un terme appliqué conventionnellement, bien que de manière pas très exacte, au type d'activité désigné dans ce volume (son livre de RI). Un SRI n'informe pas (change la connaissance de) l'utilisateur sur le sujet de sa requête. Il l'informe simplement sur l'existence (ou l'absence) de documents relatifs à sa requête et où les trouver" [Lancaster, 68].

L'un des problèmes majeurs des SRI est que la requête ou bien une partie de la requête doit être contenue dans les documents sélectionnés lors de l'opération de recherche. Cependant, ce n'est pas le cas, le système peut rater plusieurs documents pertinents pour une requête donnée si ces documents ratés contiennent des synonymes de mot en question et en même temps peut restituer des documents non pertinents à cause de présence des mots de la requête dans ces derniers.

Nous pouvons constater ici un fossé entre le but de la RI et la méthode qui la réalise: l'objectif de la RI est de retrouver des documents ayant une certaine signification (sens); alors qu'elle est implémentée de façon à ce qu'elle cherche des documents contenant les mêmes mots que ceux de la requête.

Surmonter ces limites est l'objet de plusieurs projets de recherche récents. C'est le cas notamment de l'approche de RI dite "basée concepts". Les travaux menés dans le cadre de notre thèse s'inscrivent dans cet axe. Plus précisément, il s'agit de trouver des modèles de représentation des documents et des requêtes en utilisant le réseau conceptuel d'une ontologie comme espace de représentation. Etant donné que les concepts sont des entités abstraites, les représenter est en soi un problème. Depuis la fin des années 1990, les ontologies offrent cet espace conceptuel sur lequel ces systèmes s'appuient pour saisir une partie de la sémantique présente dans les documents et les requêtes. Cette sémantique vient de l'utilisation des représentants des concepts (termes) de l'ontologie comme vocabulaire de référence qui englobe aussi bien le vocabulaire de l'utilisateur que celui que l'auteur a utilisé pour rédiger son document. Ceci permet, à l'utilisateur qui exprime un besoin en information et à l'auteur du document, de "parler le même langage".

Un SRI basé-concepts se caractérise par la notion d'espace conceptuel dans lequel les documents et les requêtes sont représentés par opposition à l'espace mots simples qu'on trouve dans les modèles classiques [Baeza-Yates et Ribeiro-Neto, 99]. Ces SRI de nouvelle génération sont prometteurs dans la mesure où ils passent du niveau symbole ("chaînes de

caractères") au niveau conceptuel. Ce qui leur permettrait de s'affranchir (ou du moins de réduire considérablement) des contraintes morphologiques, de la synonymie et de la polysémie qui sont longtemps connus en RI comme génératrices de bruit et de silence.

Une étape préalable et cruciale dans ces systèmes, est la phase d'identification et d'extraction des concepts, qui reste encore un problème ouvert. En effet, il s'agit de s'assurer dans ce cas, que tous les concepts du document (requête) sont reconnus et uniquement ceux-là. Dans la majorité des travaux décrits dans la littérature, le processus d'assimilation ou de détection de concepts (Concept Mapping) est un processus semi-automatique. Une phase supplémentaire est souvent nécessaire pour corriger les concepts trouvés par un algorithme de désambiguïsation. Ceci est dû notamment aux problèmes de synonymie et de polysémie présents dans le langage naturel.

Comme les SRI pour les autres longues, les SRI pour les textes arabes n'échappent pas aux problèmes liés à la manipulation des mots clés au lieu des concepts. Afin de résoudre ces problèmes et dans le but d'intégrer la notion de concept dans ces derniers, nous avons besoin de développer un système capable de filtrer les documents qui ne sont pas directement liés au mot désiré par l'utilisateur et présenter seulement les résultats qui correspondent à son intérêt. Ce système doit être basé sur les ontologies pour représenter aussi bien l'information (souvent des documents textuels) que le besoin en information de l'utilisateur (requête). Dans notre thèse nous envisageons de répondre aux questions :

Est-ce que l'intégration des ontologies dans les SRI pour les textes arabes apporte une valeur ajoutée pour ces systèmes ?

Dans quelle partie du SRI, l'intégration de l'ontologie donne des meilleurs résultats ?

L'axe qui a été suivi dans le cadre de notre thèse, concerne donc l'utilisation des ontologies pour une représentation sémantique de l'information en RI. Nous avons travaillé sur l'intégration de l'ontologie WordNet Arabe (AWN) dans un SRI dans la phase de représentation des documents et des requêtes par les sens extraits depuis AWN. L'Indexation Sémantique (IS) est un processus crucial dans l'opération de recherche. La qualité de l'IS dépend de la précision des techniques de mapping et de désambiguïsation utilisées pour sélectionner les concepts représentatifs des documents et requêtes. La qualité d'une RI sémantique dépend outre de la qualité de l'IS, de la qualité de la fonction d'appariement utilisée pour comparer les représentations sémantiques des documents et requêtes et calculer le degré de correspondance entre leurs représentations respectives.

Cette thèse est composée de quatre chapitres :

- Le premier chapitre présente les notions de base et les principaux concepts liés au domaine abordé dans cette thèse, à savoir, le domaine de RI. Il se focalise particulièrement sur la description des SRI et les ontologies. Il présente l'architecture générale d'un SRI telle qu'elle est admise actuellement ainsi qu'un aperçu sur les principaux modèles de recherche existants dans la littérature, il aborde aussi les différentes mesures d'évaluation de pertinence.

Le chapitre décrit aussi brièvement le concept d'ontologie : les circonstances de leur apparition, différentes définitions et les principales ontologies implémentées.

- Le deuxième chapitre décrit l'aspect utilisation des ontologies dans le domaine de la RI. Il présente aussi les différentes méthodes de désambigüisation des termes. Il montre les méthodes de désambigüisation par l'utilisation des ressources sémantiques telles que les dictionnaires, les ontologies et les corpus linguistiques.

- Le troisième chapitre propose un état de l'art sur la recherche sémantique pour les textes arabes. Ce chapitre commence par une brève description des particularités de la langue arabe. Ensuite, il aborde la problématique de la langue arabe et la RI. Finalement, les différents travaux existants qui portent sur l'utilisation des ontologies dans les SRI arabes sont ainsi discutés.

- Le quatrième chapitre présente notre contribution. Il concerne l'intégration des ontologies dans un SRI pour les textes arabes. Nous présentons différents outils implémentés. Les différents résultats correspondants à différentes expérimentations sont ensuite rapportés et commentés, pour enfin terminer par une conclusion sur l'apport des ontologies dans l'IS des informations et les perspectives de leur utilisation pour l'ensemble de la RI.

Table des matières

Chapitre 1 : Les Systèmes de Recherche d'Information & les Ontologies

1. Introduction

Aujourd'hui, la quantité d'information diffusée par nos médias est grande et de plus en plus énormément vaste, et par conséquent, la tâche qui consiste à trouver une information pertinente par un simple utilisateur devient très difficile de façon manuelle.

L'informatique a permis le développement d'outils pour faciliter la RI. Ces outils s'appellent SRI. Ces systèmes prouvent leurs efficacités au début de leur mise en œuvre par rapport à la recherche manuelle, mais avec le temps ils sont devenus incapables de répondre au besoin de l'utilisateur.

Avec l'avènement des ontologies, de nouvelles générations de SRI sont apparus. Ces derniers se basent sur la notion de concept pour représenter l'information.

Ce chapitre fait le survol des deux technologies à savoir les SRI et les ontologies.

2. Les Systèmes de Recherche d'Information

Le monde assiste depuis ces dernières décennies, a une production massive d'informations dans tous les domaines d'intérêts. L'objectif principal de la RI est de mettre en œuvre un processus de comparaison entre le besoin de l'utilisateur et les documents d'une collection dans le but de retrouver ceux qui sont pertinents.

Le terme " Recherche d'Information " (Information Retrieval) a été introduit par Calvin N. Mooers en 1948 pour la première fois quand il travaillait sur son mémoire de maitrise [Mooers, 48]. La première conférence dédiée à ce thème -International Conference on Scientific Information - s'est tenue en 1958 à Washington. On y comptait les pionniers du domaine, notamment, Cyril Cleverdon, Brian Campbell Vickery, Peter Luhn, etc. [Karbasi, 07]

Les SRI, servent d'interface entre une source (collection) contenant des quantités considérables de documents et des utilisateurs cherchant, via des requêtes, des informations susceptibles de se trouver dans cette collection. Les SRI intègrent un ensemble de techniques permettant de sélectionner ces informations. Elles peuvent être résumées en trois fonctions, qui sont le stockage de l'information, l'organisation de ces informations et la recherche et la restitution d'informations en réponse à des requêtes utilisateurs.

2.1. Les étapes d'un processus de recherche d'information

Un SRI intègre un ensemble de techniques dédiées à sélectionner dans une collection de documents ceux comportant des informations répondant au besoin de l'utilisateur qui est exprimé par une requête.

Un SRI se compose de trois fonctions principales représentées schématiquement par le processus en U de RI [Belkin et al., 92][Hlaoua, 07] (voir la figure 1.1).

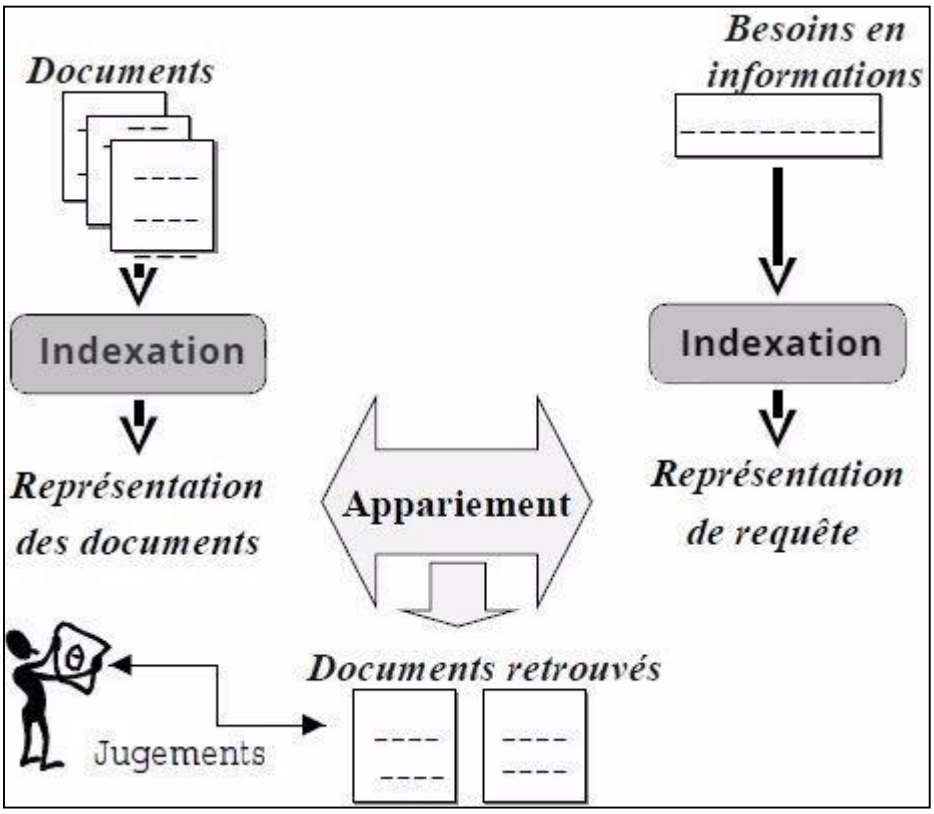


Figure 1.1 : Le processus en U de la RI

Du côté gauche, on a l'information accessible dans le système. Elle représente en général les collections de documents ou de sous collections de documents traitant d'un même domaine ou des domaines proches. Du côté droit, on a le besoin en information exprimé par la requête

de l'utilisateur. Enfin, pour combiner les deux cotés ; le système propose un certain nombre de traitement pour trouver les informations pertinentes. Ces traitements s'appuient sur un certain nombre de modèles permettant ainsi de sélectionner des informations pertinentes en réponses à une requête utilisateur. Il s'agit principalement du processus de représentation (indexation) et du processus de recherche.

Ces trois principales fonctions peuvent utiliser une ressource externe qui peut être une ontologie, une hiérarchie de concept ou le vocabulaire contrôlé d'un thesaurus. Dans ce cas, on parle d'indexation, d'appariement et de reformulation de requêtes guidés par une ontologie. Nous situons nos travaux de recherche dans ce cadre-là.

2.1.1. Processus de représentation (Indexation)

Ce processus peut être considéré comme une étape de préparation à la recherche. Il a pour rôle d'extraire des termes donnant une représentation détaillée (appelé aussi représentation paramétrique) du contenu sémantique d'une requête ou document. Ce processus est appelé indexation, il permet donc de passer d'une description brute d'un document ou d'une requête vers une description structurée. Sa qualité dépend en partie de la qualité des réponses du système. Les descripteurs reconnus par le SRI sont regroupés dans une structure appelée dictionnaire constituant le langage d'indexation. Ce langage est divisé en deux catégories :

- Le langage contrôlé : dans cette catégorie, l'intervention de l'expert est très importante pour le choix de descripteurs. L'indexation est alors dirigée de façon manuelle ou semi-automatique. Un risque de confusion peut apparaitre lorsque les termes des utilisateurs et le vocabulaire des experts s'opposent.
- Le langage libre : dans ce contexte, l'extraction des indexes se fait de manière automatique, et par conséquent le taux d'indexation est très élevé et le risque qui apparait maintenant réside dans les descripteurs non significatifs.

L'opération d'indexation peut se dérouler en trois modes différents [Kompaoré, 08] [Tamine, 00] [Nassr, 02] :

2.1.1.1. Indexation manuelle (L'indexation humaine)

Ce genre d'indexation est guidé par un spécialiste du domaine. Même s'il existe une différence entre deux spécialistes ou un spécialiste et lui-même dans le choix de l'index, cette méthode a un résultat toujours fiable. Donc ce type permet d'assurer une meilleure correspondance entre les documents et les descripteurs choisis par les indexeurs. En effet, les

spécialistes d'un domaine choisissent les meilleurs termes pour indexer les documents. Enfin, il faut noter que l'indexation manuelle est couteuse en temps.

2.1.1.2. Indexation semi-automatique

Dans ce contexte, le processus d'indexation se déroule en deux phases. Dans la première phase, les index sont extraits automatiquement et sont ensuite transmis aux spécialistes du domaine pour les valider à l'aide d'un thesaurus ou une base terminologique. Cette façon de faire est appelée aussi « indexation supervisée ».

2.1.1.3. Indexation automatique

L'indexation automatique comporte seulement les procédures automatiques sans recourir à l'intervention de l'homme. Cette méthode d'indexation est actuellement la méthode la plus répandue, elle passe par deux étapes :

- La détermination des indexes : l'extraction des indexes passe par un anti-dictionnaire pour supprimer les mots de liaisons et les mots vides ainsi que la suppression des variations des mots. Toutes ces opérations sont déterminées par une analyse morphosyntaxique.
- La pondération des indexes : la détermination des poids va différencier entre les termes par leurs importances dans le document ou la requête.

L'indexation automatique peut se faire selon deux approches : statistique et/ou linguistique. L'approche statistique se base sur la distribution statistique des termes dans le document. L'approche linguistique se base sur les techniques de traitement du langage naturel, telles que l'analyse lexicale, syntaxique et sémantique. De manière générale, l'indexation automatique est réalisée selon les étapes suivantes [Hlaoua, 07]:

- **Analyse lexicale :** L'analyse lexicale (tokenization en anglais), ce processus permet de faire le passage du texte d'un document à un ensemble de termes. Un terme est un groupe de caractères constituant un mot significatif. L'analyse lexicale permet de reconnaitre les espaces de séparation des mots, des chiffres, les ponctuations, etc.
- **L'élimination des mots vides :** Un des problèmes de l'indexation consiste à extraire les termes significatifs et éliminer les mots vides (pronoms personnels, prépositions, ...). Les mots vides peuvent aussi être des mots athématiques (les mots qui peuvent se retrouver dans n'importe quel document parce qu'ils exposent le sujet mais ne le traitent pas, comme par exemple « contenir », « appartenir »).

On distingue deux techniques pour éliminer les mots vides :

- ✓ Utilisation d'une liste de mots vides (appelée aussi anti-dictionnaire, stoplist en anglais),
- ✓ Elimination des mots dépassant un certain nombre d'occurrences dans la collection.

- **Lemmatisation :** Un mot donné peut avoir différentes formes dans un texte, par exemple "مكتبة" ،"مكاتب"،"كتاب". Pour indexer ces différentes variations des mots on utilise la technique de racinisation.

2.1.2. Pondération des termes

La pondération permet d'attribuer un poids pour un terme d'indexation afin de représenter l'importance de ce terme dans le document respectivement dans la requête [Kompaoré, 08] [Karbasi, 07].

La plupart des techniques de pondération sont basées sur les facteurs TF et IDF :

- TF (*Term Frequency*) : mesure l'importance d'un terme dans un document. Elle est utilisée pour déterminer la pondération locale. Cette mesure est proportionnelle à la fréquence du terme dans le document. Elle peut être utilisée telle quelle ou selon plusieurs déclinaisons (log(TF), présence/absence,...)
- IDF (*Inverse of Document Frequency*) : ce facteur mesure l'importance d'un terme dans toute la collection. Cette mesure est utilisée pour désigner la pondération globale. Un terme qui apparait souvent dans la base documentaire ne doit pas avoir le même impact qu'un terme moins fréquent. Il est généralement exprimé comme suit : $log(\frac{N}{df})$, où *df* est le nombre de documents contenant le terme et *N* est le nombre total de documents de la base documentaire.

La mesure TF * IDF est une bonne approximation de l'importance d'un terme dans un document. Cette mesure a eu un succès limité dans les corpus de tailles très variables.

L'inconvénient de la mesure TF * IDF est quelle ne tient pas compte de la longueur de document, en effet, un terme dans un document long a plus de chances d'apparaître plusieurs fois qu'un autre terme dans un document court. Plus le document est long, plus les termes utilisés se répètent. La longueur des documents peut aussi induire l'utilisation d'un grand nombre de termes pour décrire un sujet. Pour pallier ce type de problèmes plusieurs travaux

comme ceux de Robertson [Robertson et Walker, 94] et Singhal [Singhal et al., 96] intègrent la taille des documents dans le calcul afin de normaliser la pondération.

2.1.3. L'appariement requête-document

Le processus d'appariement requête-document est le cœur d'un SRI. Il comprend la fonction de décision fondamentale qui permet d'associer à une requête, l'ensemble des documents pertinents à restituer. La mesure de pertinence est calculée à partir d'une fonction de similarité, notée RSV (Q, d) (Retrieval Sattus Value), « Q » étant une requête et « d » un document. Elle tient compte des poids des termes déterminés en fonction d'analyses statiques et probabilistes. De manière générale, à chaque réception d'une requête, le système crée une représentation similaire à celle des documents, puis calcule un score de correspondance entre la représentation de chaque document et celle de la requête. La correspondance peut être binaire (pertinent ou non pertinent), on parle alors d'appariement exact ou on peut mesurer un degré (similarité, probabilité) de pertinence, on parle alors d'appariement approché. Idéalement, la correspondance entre ces deux représentations déterminées par le système doit s'accorder au jugement de pertinence de l'utilisateur. Pour une requête donnée, le système retourne des documents en ordre décroissant du score de pertinence. Les documents seront jugés par l'utilisateur et son jugement sera utilisé pour améliorer la représentation de la requête ; c'est ce qu'on appelle la reformulation de requêtes dans le contexte de la RI [Baziz, 05] [Nassr, 02] [Kompaoré, 08] [Tebri, 04].

Notons que d'une façon générale, le modèle de représentation des documents et requêtes ainsi que l'appariement document-requête, permettent de caractériser et d'identifier un modèle de RI. Les principaux travaux effectués dans ce domaine ont fait l'objet de modèles basés sur les approches vectorielles [Salton, 71], probabilistes [Robertson, 77], connexionnistes [Boughanem, 00], et inférentiels [Turtle et al., 92] [Haines et al., 93].

2.1.4. La notion de pertinence

On peut distinguer deux types de pertinence : la pertinence système (ou mesure de ressemblance), et la pertinence utilisateur :

2.1.4.1. La pertinence système (ou mesure de ressemblance)

Chaque SRI doit s'appuyer sur un modèle de pertinence qui lui permet de calculer pour chaque document un score de pertinence, elle apparaît donc ici non pas comme une notion subjective mais comme une valeur numérique calculée par les SRI. Cette pertinence système a

cependant des limites car elle est estimée à partir d'un score de ressemblance entre la requête et les documents, et détermine une pertinence supposée des documents pour l'utilisateur [Kompaoré, 08].

2.1.4.2. La pertinence utilisateur

C'est une notion subjective car elle dépend du niveau de satisfaction que l'utilisateur tire de la liste de documents qui lui est restituée par le système. En effet, deux utilisateurs différents ayant soumis la même requête au SRI ne jugent pas de la même manière les réponses du système. Dans le cas où le jugement de pertinence n'est pas absolu (c'est-à dire que l'utilisateur dit si le document est pertinent ou non pertinent) mais donné par un degré de pertinence des documents, le désaccord entre plusieurs utilisateurs est nettement plus prononcé. Cela est dû au fait que les besoins sont différents et que le même besoin peut être exprimé différemment en fonction de l'utilisateur. De plus, l'interprétation que l'utilisateur fait des documents qu'il reçoit dépend en partie de ses connaissances personnelles et de son expérience, ainsi que du contexte dans lequel s'effectue sa recherche. La pertinence utilisateur permet à ce dernier d'exprimer sa satisfaction par rapport aux documents potentiellement pertinents, que le système lui restitue [Kompaoré, 08].

2.1.5. Reformulation de Requêtes

En plus des étapes de représentation et d'appariement, quelques systèmes intègrent une étape supplémentaire de reformulation de requêtes. C'est un processus permettant de générer une nouvelle requête plus adéquate à la RI dans l'environnement du SRI. Son principe est de modifier la requête de l'utilisateur par ajout de termes significatifs et/ou ré-estimation de leur poids. Le processus de la reformulation (voir la figure 1.2) peut se faire de deux manières :

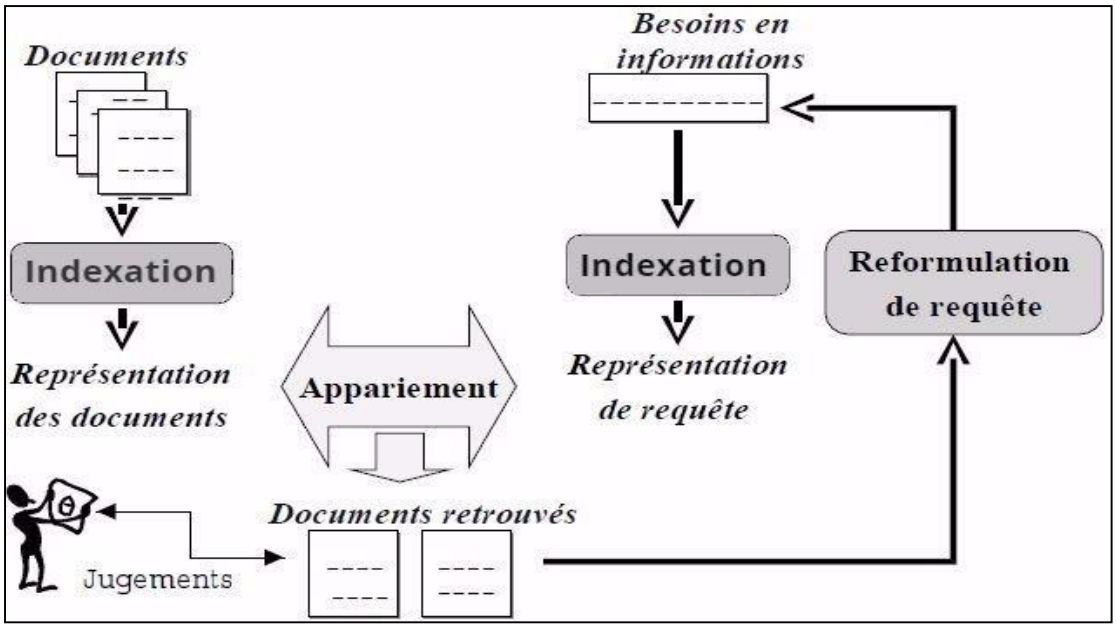


Figure 1.2 : Le processus de reformulation dans un SRI

2.1.5.1. Reformulation directe

Elle consiste à ajouter de nouveaux termes à la requête initiale. Cette modification est réalisée soit grâce aux liens de cooccurrence entre les termes ou bien en s'appuyant sur un dictionnaire dans le cadre du multilingue ou encore, comme ce qui est l'objet de notre thèse, sur une ontologie (exemple : WordNet) capable d'enrichir avec des termes dérivant de relations sémantiques telles que synonymie, spécialisation/généralisation et composition. On parle alors de reformulation de requêtes basée sur les concepts (Concept-based Query Reformulation).

2.1.5.2. Reformulation indirecte

Dans cette approche la requête est modifiée en tenant compte d'une liste de documents déjà jugés sélectionnés. Ce processus est appelé réinjection de la pertinence (relevance feed-back) si le processus est supervisé et de pseudo réinjection de pertinence si le processus est automatique. Cette méthode a un double avantage : une simplicité d'exécution pour l'utilisateur qui ne s'occupe pas des détails de la reformulation, et un meilleur contrôle du processus de recherche en augmentant le poids des termes importants et en diminuant celui des termes non importants.

2.2. Les modèles de Recherche d'Information

La fonction principale pour un SRI est de maximiser la pertinence des résultats trouvés. Un modèle de RI fournit une formalisation au processus de RI. Il doit accomplir plusieurs rôles dont le plus important est de fournir un cadre théorique pour la modélisation de la mesure de pertinence.

Dans ce qui suit, nous décrivons brièvement les modèles de RI, et quelques modèles dérivés ou inspirés à partir de ces classes.

2.2.1. Les modèles booléens

2.2.1.1. Le modèle booléen

Le modèle booléen est le premier modèle inventé par salton [Salton, 71], ce modèle est basé sur la théorie des ensembles et l'algèbre de Boole. Dans ce modèle, les documents et les requêtes sont représentés par des ensembles de mots clés. Le processus de recherche mis en œuvre via ce type de modèle, consiste à effectuer des opérations logiques utilisant des connecteurs « ET », « OU », « NON » sur les ensembles de documents, définis par l'occurrence ou l'absence de termes d'indexation. Le modèle booléen utilise le mode d'appariement exact (0 ou 1), il ne restitue que les documents répondant exactement aux termes de la requête.

2.2.1.2. Le modèle booléen étendu

Ce modèle est une extension du modèle booléen introduit en 1983 par Fox et Salton [Fox, 83]. Appelé aussi modèle P_Norm (tel que l'opérateur Lp- Norm est défini pour la mesure de pertinence requête-document). Ce modèle complète le modèle de base en intégrant des poids dans l'expression de la requête et des documents afin de mesurer le score de pertinence. Ceci a pour conséquence la sélection de documents sur la base d'un appariement rapproché (fonction d'ordre) et non exact.

2.2.1.3. Le modèle des ensembles flous

Une autre extension du modèle booléen est inspirée de la théorie des ensembles flous a été proposée par salton [Salton, 89]. Ce modèle est basé sur l'appartenance probable et non certaine d'un élément à un ensemble. L'idée de base est de traiter les descripteurs des documents et requêtes comme étant des ensembles flous. Cette extension vise également à

tenir compte de la pondération des termes dans les documents. Un poids d'un terme exprime son degré d'appartenance à un ensemble.

2.2.2. Les modèles vectoriels

2.2.2.1. Le modèle vectoriel de base

La première idée de représenter les documents et les requêtes sous forme de vecteurs de termes pondérés a été proposée par Luhn [Luhn, 57] à la fin des années cinquante. Elle a été ensuite développée par Gérard Salton et son équipe [Salton, 71] [Salton, 83] dans leur projet SMART (Salton's Magical Automatic Retriever of Text). L'idée de base du modèle vectoriel est d'utiliser une représentation géométrique pour classer les documents par ordre de pertinence par rapport à une requête. Dans ce modèle les documents et les requêtes sont engendrés par les termes d'indexation représentés par des vecteurs [Salton, 83].

2.2.2.2. Le modèle vectoriel généralisé

Dans ce modèle chaque terme est représenté par un vecteur dans un espace vectoriel dont les axes sont orthogonaux par construction : les axes sont en fait les produits logiques des termes d'indexation. Un document est représenté par la moyenne des vecteurs représentant les termes qu'il contient. Lors du calcul de pertinence ce modèle combine entre les poids des documents et le facteur de corrélation entre les vecteurs.

2.2.2.3. Le modèle LSI (Latent Semantic Indexing)

Selon [Dumais, 94], le modèle LSI est basé sur une représentation conceptuelle où les effets dus à la variation d'usage des termes dans la collection sont nettement atténués, les racines par exemple, et les documents qui partagent des termes proches doit avoir des représentations proches dans l'espace défini par le modèle. Par conséquent le système permet de sélectionner des documents même s'ils ne contiennent aucun terme de la requête.

2.2.2.4. Le modèle connexionniste

Les SRI basés sur le modèle connexionniste sont fondés sur le réseau de neurones [Kwork, 89], [Boughanem, 92] [Mothe, 94]. Ce réseau est construit à partir des représentations des contenues des requêtes et des documents sous forme de couches répartis généralement dans ce sens : requête - termes - documents [Kwork, 95]. Le mécanisme de RI basé sur ce modèle est fondé sur le principe de d'apprentissage ce qui permet aux SRI de devenir adaptatifs.

2.2.3. Le modèle probabiliste (Probabilistic Model)

Le modèle de recherche probabiliste utilise un modèle mathématique fondé sur la théorie de la probabilité [Salton, 83]. L'idée générale de cette approche est d'implémenter les notions de la théorie de probabilité sur les SRI. Le principe de base du modèle probabiliste consiste à présenter les résultats de recherche d'un SRI dans un ordre basé sur la probabilité de pertinence d'un document vis-à-vis d'une requête.

2.2.3.1. Le modèle BIR

L'idée de base du modèle BIR (Binary Independence Retrieval) est de représenter les termes des documents par des valeurs binaires (1 si le terme apparait dans un document, 0 sinon). Le jugement de la pertinence d'un document par rapport à une requête faite par le calcule de probabilité qui est basé sur le théorème de Bayes

2.2.3.2. Le modèle de réseau infèrentiel bayésien

Ce modèle été inventé par par Turtle [Turtle et Croft, 91]. Les probabilités conditionnelles de chaque nœud sont calculées par propagation des liens de corrélation entre eux. Ce modèle vise à considérer la dépendance entre les termes mais engendre une complexité de calcul importante.

2.2.3.3. Le modèle du langage

L'objectif d'un modèle de langue est de capter les régularités linguistiques d'une langue en observant la distribution des mots ou les successions de mots dans la langue donnée. Le modèle de langue désigne une fonction de probabilité qui assigne à chaque séquence de mots une probabilité. En RI, les modèles de langues déterminent la probabilité pour que la requête puisse être générée par le modèle de langue du document. Les initiateurs de ce modèle Ponte et Croft [Ponte et al., 98] disent que le calcul de cette probabilité repose sur l'hypothèse suivante : un utilisateur en interaction avec un SRI fournit une requête en pensant à un ou plusieurs documents qu'il souhaite retrouver. La requête est alors inférée par l'utilisateur à partir de ces documents. Un document n'est pertinent que si la requête utilisateur ressemble à celle inférée par le document.

2.3. Évaluation des SRI

Après la réalisation du système, l'étape d'évaluation intervient pour mesurer la fiabilité et la satisfaction d'un SRI. En effet elle permet de caractériser le modèle et de fournir des

éléments de comparaison entre les modèles. Les SRI qui sont aujourd'hui destinés à des utilisateurs non spécialistes permettent une grande richesse d'exploration en facilitant la consultation directe des documents, pour cela une étude d'évaluation est nécessaire dans le sens où elle permet de contrôler et d'évaluer les opérations et la performance du système. Ce dernier doit respecter les deux conditions suivantes :

1. retrouver tous les documents pertinents,
2. rejeter tous les documents non pertinents.

Selon [Tebri, 04] [Karbasi, 07] l'évaluation d'un SRI peut être appréhendée selon deux aspects : un aspect efficience qui est lié au rendement (rapidité et/ou quantité de ressources utilisées), il dépend de l'évaluation cognitive de l'utilisateur, tels que la facilité d'utilisation du système, rapidité d'accès, temps de réponse à une requête, présentation des résultats, nombre d'entrée-sortie, etc. et un aspect efficacité qui est lié à la qualité du résultat, et qui concerne la capacité du système à sélectionner le maximum de documents pertinents et un minimum de documents non pertinents. Nous nous intéressons dans cette section à présenter l'aspect efficacité qui est souvent mesuré par deux paramètres rappel/précision. L'évaluation de l'efficacité d'un SRI repose en général sur les trois principaux éléments suivants :

– une collection de document de test,
– un ensemble de requête,
– une liste de documents pertinents pour chaque requête, produite par des experts.

A partir de ces trois éléments, nous pouvons mesurer les taux de performance des SRI par différentes mesures d'évaluation que nous décrivons ci-dessous :

2.3.1. Les mesures de Rappel/Précision

Les mesures de précision/rappel sont obtenues en partitionnant l'ensemble des documents restitues par le SRI en deux catégories : les documents pertinents et les documents non pertinents.

– Taux de précision : La précision mesure la capacité du système de rejeter tous les documents non pertinents à une requête. Il est donne le rapport entre l'ensemble des documents sélectionnés pertinents et l'ensemble des documents sélectionnés [Baziz, 05].
– Taux de rappel : Le rappel mesure la capacité du système à retrouver tous les documents pertinents répondants à une requête. Il est donné par le rapport entre les documents retrouvés pertinents et l'ensemble des documents pertinents de la base [Baziz, 05].

Les taux de précision et de rappel sont donnés par les formulations suivantes :

$$précision = \frac{R_+}{M} \qquad\qquad rappel = \frac{R_+}{R}$$

Où: *R*: le nombre total de documents pertinents dans la collection,

M : le nombre de documents sélectionnés,

R+ : le nombre de documents pertinents sélectionnés.

La Figure 1.3 illustre la précision et le rappel du système. La zone de haute précision représente le cas où le système retourne peu de documents non pertinents, alors que la zone de haut rappel concerne le cas où seule une petite partie des documents pertinents, est omise par le système.

Le taux de rappel et le taux de précision évaluent respectivement les notions de bruit[1] et de silence[2] documentaire qui constituent les deux premiers indicateurs de performance d'un SRI.

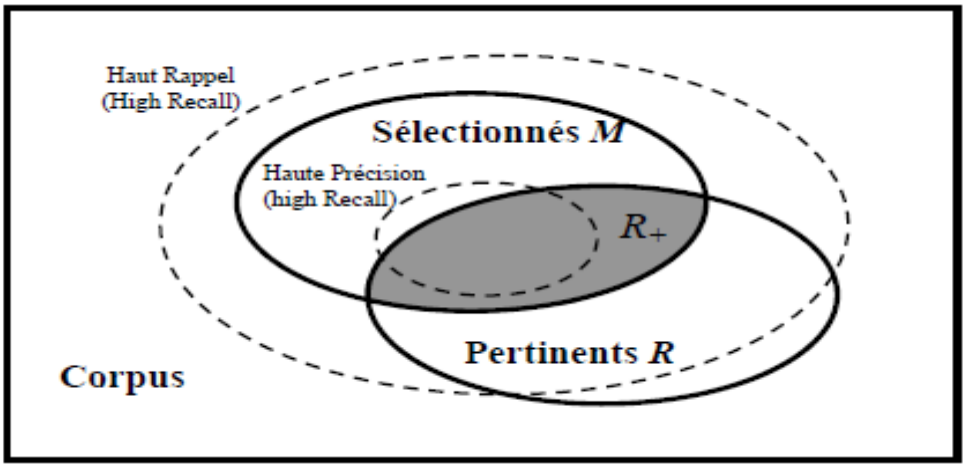


Figure 1.3 : Représentation schématique des zones de précision et de rappel [Baziz, 05]

La notion de silence et bruit présentent respectivement le taux de documents pertinents non sélectionnés et le taux de documents non pertinents sélectionnés. Ils sont donnés par les deux formules suivants : le Silence = 1 - Rappel et le Bruit = 1 - Précision.

[1] Le bruit fait référence aux documents non pertinents retrouvés par le système.

[2] Le silence fait référence aux documents pertinents mais qui ne sont pas retrouvés par le système.

2.3.2. La courbe de Rappel / Précision

Une façon d'évaluer un système est de tracer une courbe de précision-rappel (voir figure 1.4). Lorsque le (rappel=1) et la (précision=1) on parle d'un SRI idéal. Cependant, le rappel et la précision sont deux mesures qui varient généralement en sens inverse.

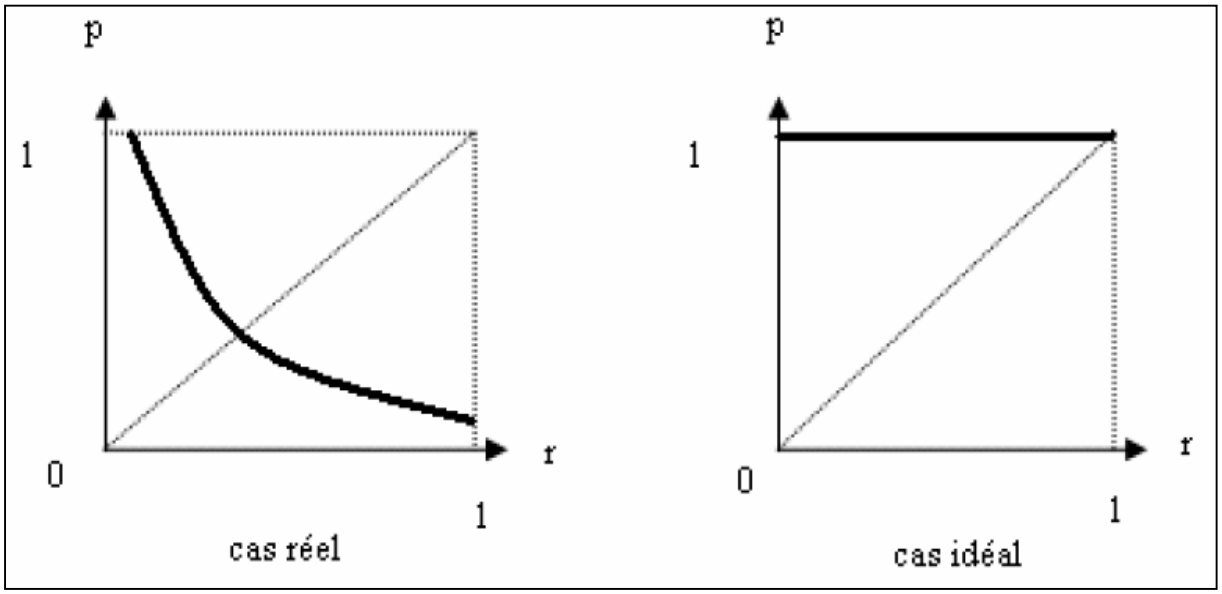


Figure 1.4 : Courbe de rappel/précision [Karbasi, 07].

Un système dont la courbe dépasse (c'est-à -dire qu'elle se situe en haut à droite de) celle d'un autre est considérée comme un meilleur système. Il arrive parfois que les deux courbes se croisent. Dans ce cas, il est difficile de dire quel système est le meilleur.

3. Les ontologies

Dès l'annonciation de naissance des ontologies et les chercheurs de domaine informatique sont en concurrence pour les intégrer dans les déférents domaines comme par exemple l'ingénierie des connaissances, le traitement du langage naturel (NLP), les systèmes d'informations coopératives, l'intégration intelligente d'information, la gestion des connaissances et la RI. L'intégration des ontologies dans les SRI permet de doter ces systèmes par un peu de sémantique et ainsi diminue la divergence entre le besoin de l'utilisateur et les réponses système.

3.1. Définitions des ontologies

« Ontologie » est un mot de l'informatique issu du domaine philosophique apparut au début des années 90 [Gruber, 93]. Les ontologies sont introduites au champ de l'Intelligence Artificielle (IA) et de la RI afin de représenter les connaissances, partager l'information et faciliter la communication.

Voici quelques définitions des ontologies :

Définition 1 : « Une ontologie peut prendre différentes formes, mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification. Cette dernière inclut des définitions et une indication de la façon dont les concepts sont reliés entre eux, les liens imposent collectivement une structure sur le domaine et contraignent les interprétations possibles des termes » [Baziz, 02].

Une telle caractérisation rend compte d'objets divers tels des glossaires, des terminologies, des thesaurus et des ontologies (au sens strict), mis en œuvre par différents professionnels (ingénieurs de la connaissance, bibliothécaires, traducteurs) et se distinguent suivant que l'accent est mis sur les termes ou leur signification. Dans le contexte de l'intelligence artificielle on peut trouver d'autres définitions :

Définition 2 : « une ontologie définit les termes et les relations de base du vocabulaire d'un domaine ainsi que les règles qui indiquent comment combiner les termes et les relations de façon à pouvoir étendre le vocabulaire » [Baziz, 02].

Cette définition explique la façon d'élaborer une ontologie en nous offrant des directives relativement floues : repérer les termes de base et les relations entre les termes, identifier les règles servant à les combiner, fournir des définitions de ces termes et de ces relations. Notons que d'après cette définition, une ontologie inclut non seulement les termes qui y sont explicitement définis, mais aussi les termes qui peuvent être créés par déduction en utilisant les règles. En 93, Gruber [Gruber, 93] formule la définition suivante :

Définition 3 : « une ontologie est une spécification explicite d'une conceptualisation ».

Cette définition est la plus citée et restera la plus juste. Elle signifie que la construction d'ontologie intervient après une étape de conceptualisation. La définition de Guarino et Giaretta, « Une ontologie est une spécification rendant partiellement compte d'une conceptualisation ». La conceptualisation étant spécifiée parfois de manière très précise, une

théorie logique ne peut pas toujours en rendre compte de façon exacte : elle ne peut assumer la richesse interprétative du domaine conceptualisé dans une ontologie et ne le fait donc que partiellement. Pour Borst, en modifiant légèrement la définition de Gruber : « l'ontologie est une spécification formelle d'une conceptualisation partagée » [Baziz, 02].

Définition 4 : « Une ontologie est un ensemble de termes structurés de façon hiérarchique, conçu afin de décrire un domaine et qui peut servir de charpente à une base de connaissances » [Baziz, 02].

Cette définition se base sur le fait qu'ils construisent des ontologies de connaissances spécifiques à des domaines d'expertise en identifiant les termes significatifs d'un certain domaine de l'ontologie Sensus (qui inclut plus de 50 000 termes).

Bernaras et ses collègues construisent une ontologie différemment, en partant d'une base de connaissances qui sera raffinée et enrichie de nouvelles définitions si de nouvelles applications sont créés. Une définition a été proposée dans ce sens par [Gomez, 99]:

Définition 5 : « une ontologie fournit les moyens de décrire de façon explicite la conceptualisation des connaissances représentées dans une base de connaissances. » La Linguistique est aussi concernée par la question des ontologies dans la mesure où les données dont on dispose pour élaborer les ontologies consistent en des expressions linguistiques de connaissances [Baziz, 02].

La caractérisation du sens de ces expressions conduit à déterminer des signifiés contextuels, dépendants des contextes (documents) où les expressions apparaissent. Ces signifiés contextuels doivent alors être normés, ce qui revient à fixer une signification pour un contexte de référence, celui de la tâche (application) pour laquelle l'ontologie est élaborée [Bachimont, 00].

Pour conclure cette section, nous pouvons donc affirmer que les définitions du terme ontologie traitent les connaissances, leurs définitions et leurs manipulations.

3.2. Composants des ontologies

Les concepts : sont utilisés dans leur sens large. Ils peuvent être abstraits ou concrets, élémentaires (électrons) ou composés (atomes), réel ou fictifs. En résumé, un concept peut être tout ce qui peut être évoqué : description d'une tâche, d'une fonction, d'une action, d'une stratégie ou d'un processus de raisonnement [Baziz, 02].

Les relations : représentent un type d'interaction entre les notions d'un domaine. Elles sont formellement définies comme tout sous-ensemble d'un produit de n ensembles, c'est à dire R : C1 x C2 x … x Cn. Comme exemple de relation binaire, on peut citer sous-classe-de ou encore connecté-à [Baziz, 02].

Les fonctions : sont des cas particuliers de relations dans lesquelles le nième élément de la relation est défini de manière unique à partir des n-1 premiers. Formellement, les fonctions sont définies ainsi : F : C1 x C2 x … x Cn-1 -> Cn. Comme exemple de fonctions binaires, nous avons la fonction mère de [Baziz, 02].

- *Les axiomes* : pour structurer des phrases qui sont toujours vraies.
- *Les instances* : elles sont utilisées pour représenter des éléments.

3.3. Les principaux types d'ontologies

Le domaine des ontologies est très vaste et ses utilisations comprennent plusieurs champs. De manière générale, on identifie les types suivants : Les ontologies de représentation des connaissances, les méta-ontologies, les ontologies de domaine, les ontologies de tâches, les ontologies de domaine-tâche, les ontologies d'application ainsi que les ontologies interactives (voir la figure 1.5) :

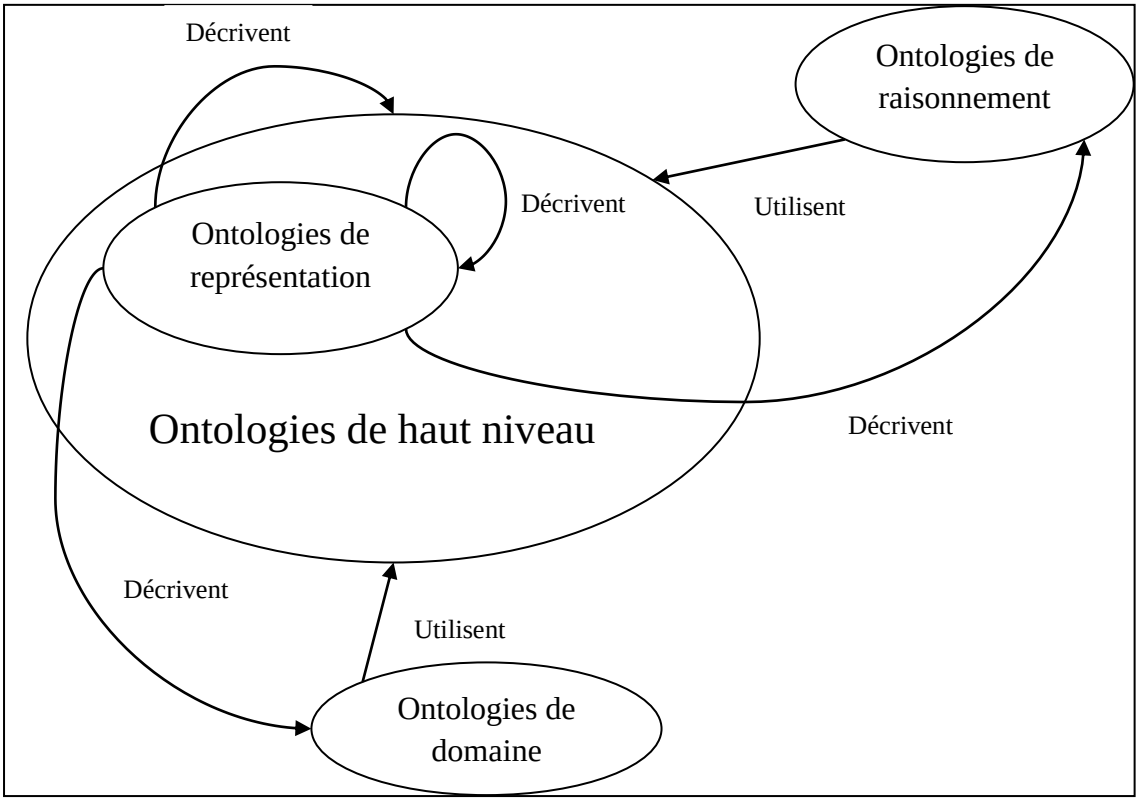


Figure 1.5 : les différents types d'ontologies [Fürst, 04]

Selon [Baziz, 02] :

- *Les ontologies de représentation de connaissance* : regroupent les primitives de représentation utilisées afin de formaliser les connaissances. L'exemple le plus représentatif de ce type d'ontologie est la Frame-Ontology [Gruber, 93], qui rassemble les primitives de représentation (classes, instances, cases, facettes, etc.) utilisées dans les langages à base de frame.
- *Les ontologies générales/communes* : incluent le vocabulaire lié aux objets, aux événements, au temps, à l'espace, à la causalité, au comportement et à la fonction.
- *Les méta-ontologies* : également appelées ontologies génériques ou noyaux d'ontologies, spécifiant les processus de raisonnement appliqués aux connaissances. Ontologies représentant les connaissances génériques mises en œuvre lors de la résolution automatique de problèmes. Seules sont décrites les connaissances portant sur la façon d'utiliser d'autres connaissances. Ces ontologies sont réutilisables dans différents domaines. L'exemple le plus représentatif serait une ontologie méréologique [Borst, 97], qui inclurait le terme *partie de.*
- *Les ontologies de domaine* : sont réutilisables dans un domaine donné. Elles fournissent le vocabulaire des concepts d'un domaine (par exemple scalpel, scanné dans un domaine médical) et les relations entre ces derniers, les activités de ce domaine (par exemple anesthésie, accouché) ainsi que les théories et les principes de base de ce domaine.
- *Les ontologies de tache* : fournissent un vocabulaire systématisé des termes utilisés pour résoudre les problèmes associés à des taches qui peuvent appartenir ou non à un même domaine. Ces ontologies fournissent un ensemble de termes au moyen desquels on peut décrire au niveau générique comment résoudre un type de problème. Elles incluent des noms génériques (par exemple plan, objectif, contrainte), des verbes génériques (par exemple assigner, classer, sélectionner), des adjectifs génériques (comme assigné) et d'autre mots qui relèvent de l'établissement d'échéances.
- *Les ontologies de domaine-tâche* : Ce sont des ontologies de tâches réutilisables dans un domaine donné, mais pas dans différents domaines. Par exemple une ontologie domaine-tâche dans le domaine médical, pourrait inclure les termes liés au timing d'une intervention chirurgicale : planifier-intervention chirurgicale.
- *Les ontologies d'application* : Contiennent suffisamment de connaissances pour structurer un domaine particulier.

3.4. Les ontologies les plus connues

Cette section introduit les ontologies les plus connues en prenant en compte la typologie d'ontologies énoncée ci-dessus. Actuellement, nous trouvons sur le web plusieurs ontologies, telles que Ontolingua sur le serveur Ontolingua [Farquhar et al., 97] et WordNet [Miller, 90] à Princeton disponibles gratuitement. D'autres ontologies, telles que les ontologies de Cyc [Lenat et al., 90] sont partiellement disponibles gratuitement sur le Web. La majorité d'entre elles sont propres à des entreprises et leurs utilisations n'est pas gratuite.

Voici quelques ontologies les plus connues [Baziz, 02]:

- L'ontologie *Page* (également connue sous le nom de *Top*) et « *(Onto) 2Agent »* qui est un moteur de recherche sur Internet basé sur une ontologie et qui aide à sélectionner des ontologies.
- Les ontologies de haut niveau : fournissent des concepts généraux permettent de définir tous les termes des ontologies existantes. Parmi elles, KR Ontology, le Penman Upper Level et Cyc.
- L'*ontologie méréologique* : pourrait être l'exemple typique d'une méta-ontologie. Cette ontologie définit la relation *partie-de* et ses propriétés. Cette relation permet d'exprimer que des instruments sont formés de composants, qui peuvent eux même être constitués d'éléments plus petits.
- L'ontologie *Cyc:* est une ontologie de sens commun qui fournit une grande quantité de savoir humain élémentaire. Elle consiste en un ensemble de termes et d'affirmations liées à ces termes. Elle se décompose par ailleurs, en différentes micro-théories. Chaque micro-théorie rend compte seulement d'un point de vue important d'un domaine de connaissances. Certains domaines peuvent traiter plusieurs micro-théories qui représentent différentes perspectives et affirmations et divers niveaux de granularité et de distinction. Les *ontologies Cyc* sont implémentées dans un langage appelé Cycl.
- Le *Generalized Upper Model (Gum), WordNet* [Miller, 90] et *Sensus* [Swartout et al., 97] représentent le mieux les ontologies linguistiques. Le Generalized Upper Model est une ontologie linguistique générale, indépendante de tout domaine et de tout type de tâche. Afin de pouvoir la transférer dans différentes langues, il a été prévu que l'ontologie *Gum* inclus seulement les notions linguistiques principales et leur organisation dans toutes les langues. Elle omet ainsi les détails qui différencient les langues. Cette façon de concevoir a permis d'utiliser *Gum* pour créer des ontologies

pour des langues spécifiques, telles que l'anglais, l'allemand, l'espagnol et l'italien en rajoutant les traits sémantiques propres à chaque langue.

- WordNet est un dictionnaire informatisé développé à l'université de Princeton par un groupe dirigé par George Miller [Miller, 90]. Il est actuellement à la version 3.0, c'est une ontologie linguistique générale pour l'anglais disponible gratuitement sur Internet[3]. On peut l'utiliser « En ligne », ou télécharger la base avec une interface logicielle qui peut être intégrée aux programmes. L'ontologie WordNet peut être utilisée dans les domaines de traduction automatique et de désambigüisation de textes comme dans les projets Ontoseek [Guarino et al., 99] et Sensus. Etant donné qu'elle n'est pas spécifique à un domaine particulier, Elle est bien adaptée à la RI dans les collections de types « news » telles que Clef et TREC déjà utilisées et qui sont des références dans le domaine de la RI. Cette ontologie linguistique couvre la grande majorité des noms, verbes, adjectifs et adverbes de la langue anglaise. Les mots dans WordNet sont organisés en un ensembles de synonymes appelés « Synsets ». Chaque synset est représenté par un concept. WordNet actuelle à un vaste réseau de 155 287 mots, organisés en 117 659 synsets. Dans WordNet, le réseau de mots est exprimé par plusieurs relations sémantiques, on trouve parmi elles [Baziz, 02]:
    - Relation ***Synonymie :*** le synset (synonym set), représente un ensemble de mots qui sont interchangeables dans un contexte donné. C'est le terme générique utilisé pour désigner une classe englobant des instances de classes plus spécifiques.
    - Relation ***Hyponymie*** : c'est le terme spécifique utilisé pour désigner un membre d'une classe (relation inverse de Hypernymie). **X** est un hyponyme de **Y** si **X** est un type de (kind of) **Y**.
    - Relation ***Holonymie*** : le nom de la classe globale dont les noms méronymes font partie. Y est un holonyme de X si X est une partie de (is a part of) Y.
    - Relation ***Méronymie*** : Le nom d'une partie constituante (part of), substance de (substance of) ou membre (member of) d'une autre classe (relation inverse de l'holonymie). **X** est un méronyme de **Y** si **X** est une partie de **Y**. exemple : {*avion*} Has_meronym {{porte}, {moteur}} ; {moteur} Has_meronym {{*hélice*}, {*réacteur*}.
    - En plus de ces relations, WordNet possède quelques autres relations qui sont cependant moins utilisées en pratique. C'est le cas de la relation **antonym** pour

[3] http://www.wordnet.princeton.edu/

exprimer les sens opposés pour les synsets, de la relation **entailment** pour les verbes : Un verbe X entails (nécessite) Y si X ne peut être fait à moins que Y ne le soit ou n'ait été déjà fait. Ou encore de la relation **troponym** pour les verbes : X est un troponyme de Y si to X est pareil que to Y dans une certaine façon.

- *Sensus* est une ontologie basée sur le langage naturel, qui a pour fonction de fournir une vaste structure *conceptuelle* aux travaux menés en matière de traduction automatique. Il elle a été mise au point en rassemblant et en extrayant des données de ressources électroniques telles que : *Penman Upper Model, Ontos, WordNet* et des dictionnaires électroniques de langages naturels. Elle compte plus de 50 000 notions.
- Dans le domaine des ontologies d'ingénierie, les ontologies *EngMath* et *PhysSys* sont les plus connues. *EngMath* est une ontologie *Ontolingua* mise au point pour la modélisation mathématique en ingénierie. *PhysSys* est une ontologie d'ingénierie destinée à modéliser, simuler et concevoir des systèmes physiques : Représentation du système, comportement de processus physique, et relations mathématiques descriptives.
- Les ontologies qui représentent le mieux les ontologies dédiées à la modélisation d'entreprises sont l'Enterprise Ontologie et la Tove Ontologie. L'Enterprise Ontology est un ensemble de termes et de définitions pertinent pour les entreprises commerciales et inclut des connaissances sur les activités et les processus, les organisations, les stratégies, le marketing, etc.
- Les ontologies élaborées dans le cadre du projet Tove[4] sont l'ontologie de conception d'entreprises, l'ontologie des projets, l'ontologie-agenda, ou encore l'ontologie des services.
- L'ontologie *(KA)*[5]constitue une référence pour les ontologies dédiées à la gestion des connaissances, elle est utilisée par le *Knowledge Annotation Initiative* de la communauté d'acquisition des connaissances. Cette ontologie servira de base pour annoter les documents sur internet de la communauté d'acquisition des connaissances de façon à fournir un accès intelligent à ces documents. Des spécialistes situés dans des zones géographiques différentes travaillent ensemble à la mise au point de cette ontologie.

[4] Toronto Virtual Enterprise
[5] Knowledge Annotation

4. Conclusion

Nous avons présenté dans ce chapitre les principales notions et concepts de la RI. Nous y avons développé les principales étapes d'un processus de RI, à savoir, la représentation ou indexation des documents, la comparaison de l'information et du besoin en information. Les principaux modèles existants dans la littérature ont été également présentés, ainsi que les différentes méthodes et cadres connus d'évaluation des performances des SRI.

Les SRI sont conçus à l'origine pour retrouver les documents traitant d'un sujet donné. Or, la majorité des systèmes actuels, présentés ici, se contentent de chercher les documents qui contiennent les mêmes mots que ceux de la requête. Ceci est évidemment insuffisant dans la mesure où l'utilisateur du système ne connaît pas à priori le vocabulaire que l'auteur a utilisé dans le document. Actuellement, ces systèmes décrivent l'information par une liste de mots simples (bag of words).

Dans ce chapitre nous avons aussi vu les déférentes définitions d'ontologie ainsi que les différents types d'ontologies et les plus connues de ces derniers. Actuellement, un effort particulièrement considérable est fourni dans la recherche par des systèmes capables de retrouver des documents demandés par des utilisateurs ne connaissant pas nécessairement le vocabulaire ni la nature des documents qu'ils cherchent. Ce sont des systèmes guidés par des dictionnaires, des ontologies, des thésaurus…etc. Nous décrivons dans le chapitre suivant l'utilisation des ontologies dans le domaine de la RI.

Table des matières

Chapitre 2 : Utilisation des Ontologies pour la Recherche d'Information

1. Introduction

Après le succès d'utilisation des ontologies dans le domaine de représentation des connaissances, les chercheurs de domaine informatique sont en concurrence pour les intégrer dans les déférents domaines comme par exemple l'ingénierie des connaissances, le traitement du langage naturel (NLP), les systèmes d'informations coopératives, l'intégration intelligente d'information, la gestion des connaissances et la RI qui est le domaine qui nous intéresse dans le cadre de cette thèse. Cette intégration dans les SRI est liées principalement à la manipulation des connaissances partagées par les ontologies afin de doté ces systèmes par un peu de sémantique et ainsi diminuer la divergence entre le besoin de l'utilisateur et les réponses système.

Une fois l'ontologie choisie, la connaissance qu'elle représente peut être utilisée à différents niveaux dans le processus de RI. Elle peut aider à l'indexation des documents appelée aussi IS. Les ontologies peuvent également aider à la formulation du besoin de l'utilisateur et à l'accès aux documents. Enfin l'ontologie peut être utilisée dans le modèle lui-même pour réaliser l'appariement entre le besoin et les granules documentaires. Ces aspects sont présentés dans la section 2. L'intégration des ontologies dans ces processus est principalement basée sur l'exploitation des notions de mesures de similarités entre concepts de l'ontologie. La section 3 aborde le problème de la désambigüisation. La dernière section 4 est consacrée à l'étude de l'apport des ontologies dans les SRI.

2. Ontologies et recherche d'information

Un des enjeux actuels de la RI est de développer des systèmes capables d'intégrer plus de sémantique dans leurs traitements. L'idée est d'avoir une solution au problème de confusion entre le besoin de l'utilisateur exprimé par une requête et le domaine exprimé par une collection de documents, autrement parlé le même langage. Pour cela les ontologies interviennent afin d'améliorer la qualité des documents restitués par les SRI à partir d'une

collection. Elles sont utilisées pour représenter des descriptions partagées et plus ou moins formelles de domaines et ainsi ajouter une couche sémantique aux systèmes informatiques.

La question qui se pose à ce niveau est : Dans le domaine de la RI électronique tel qu'il est connu actuellement en utilisant des SRI, comment une ontologie peut-elle être associée au processus de RI ? Autrement, à quel niveau de SRI l'ontologie peut intervenir ?

2.1. Le choix d'une ontologie

La majorité des approches de RI cherchent à intégrer une ontologie existante déjà dans leur processus [Hearst, 97][Vallet, 05][Baziz, 05]. De façon générale, le seul critère prise en compte pour le choix de l'ontologie est le domaine de connaissance représentée dans l'ontologie qui doit couvrir le domaine traité dans le corpus. C'est le cas par exemple du système Cat-a-cone qui repose sur la hiérarchie de concepts du domaine de la médecine MESH [Hearst, 97] pour explorer une collection documentaire du même domaine, ou bien des travaux présentés dans [Baziz, 05] qui repose sur l'ontologie générale WordNet pour une tâche de RI ad-hoc sur une collection de TREC[6].

L'évaluation de la réutilisabilité d'ontologie pour la RI se place dans ce contexte. Les ontologies utiles pour la RI doivent être adaptées à la tâche de RI considérée et plus particulièrement apporter de la connaissance utile pour l'interprétation et la compréhension par le système des informations contenues dans le corpus documentaire.

Une première solution vise à construire une ontologie à partir du ou des corpus sur lesquels les tâches de RI vont être réalisées. Cette solution assure a priori l'adéquation entre l'ontologie construite, le corpus et la tâche à réaliser. Cette solution n'est pas toujours adaptée : elle est coûteuse et ne prend pas en compte l'existence de ressources qui pourraient être réutilisées. Maintenant avec l'avènement de domaine des ontologies, elles sont devenues des standards à réutiliser. Une autre solution très utilisée par la majorité des approches de RI visent à intégrer ces ontologies dans ces approches [Baziz, 05].

2.2. Principe d'utilisation des ontologies par un SRI

L'ontologie peut être utilisée dans les différentes phases d'un processus de SRI (voir la figure 2.1), ainsi elle peut être utilisée dans le système d'indexation des documents ainsi des requêtes, dans le processus de filtrage d'information et finalement dans le processus de

[6] Text Retrieval Conference, http://www.trec.nist.gov

recherche lui-même, c.à.d. au mécanisme de comparaison entre la représentation de requête et documents de la collection.

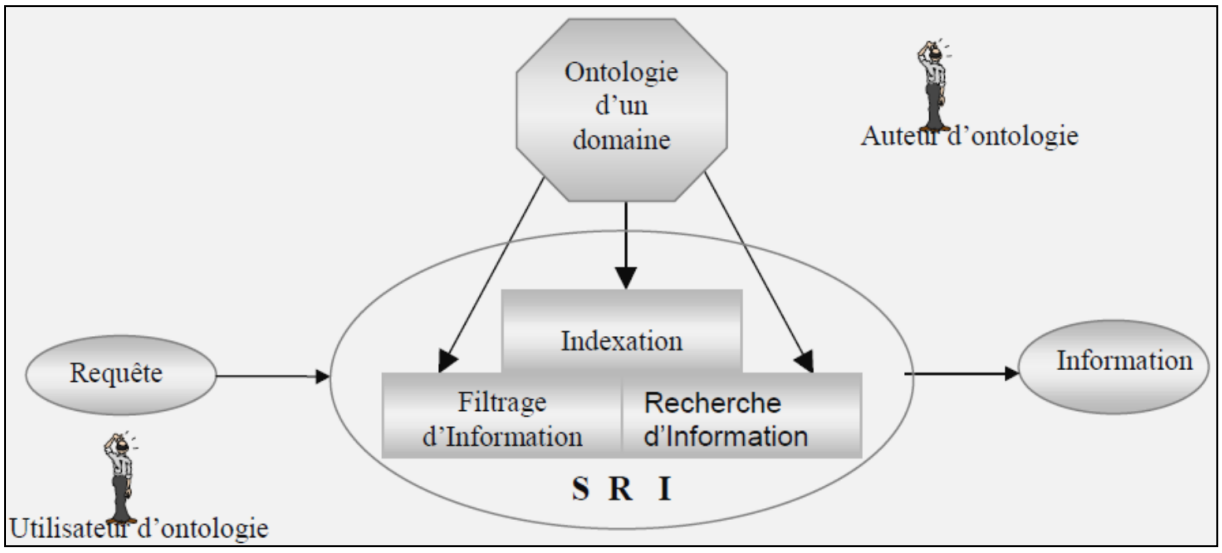


Figure 2.1 : L'ontologie greffée au processus de RI [Baziz, 02]

2.2.1. L'ontologie et la représentation des documents (Indexation)

L'indexation des documents et requêtes à l'aide des mots neutre prouve qu'elle est insuffisante et ne donne pas des bons résultats [Khan, 00]. Les chercheurs ont pensé d'ajouter un peu de sémantique dans les termes choisis comme des indexes à travers une ontologie. Ce type de traitement s'appel l'IS.

L'IS n'est possible que par l'existence et l'utilisation de ressources décrivant explicitement l'information correspondant aux objets. L'utilisation d'ontologies sous forme de hiérarchies de concepts d'ontologies légères[7] ou lourdes est le prolongement de l'utilisation dans le cadre de la RI des ressources terminologiques [Haav et Lubi, 01]. L'ontologie utilisée dans ce cas reflétant le ou les domaines de connaissance étudiés à la collection.

Dans la littérature, il existe de nombreuses définitions de l'IS. Certains auteurs différencient l'IS de l'Indexation Conceptuelle (IC) [Mihalcea et Moldovan, 00]. Pour eux, L'IC repose sur des hiérarchies de concepts ou ontologies de domaine, alors que l'IS repose sur l'utilisation d'ontologies génériques telles que WordNet. Selon [Baziz, 05] l'IC peut être vue comme une généralisation de l'IS, dans la mesure où les concepts aussi véhiculent des sens. Il sépare l'IS et conceptuelle pour deux raisons :

[7] Les ontologies dites « légères » contiennent des concepts et des relations entre concepts ainsi qu'un lexique permettant de référencer les concepts et les relations mais n'intègrent pas d'axiomes dans leur formalisation contrairement aux ontologies lourdes.

- (1) la première est due au fait que l'IS en RI se base historiquement sur les techniques de désambiguïsation pour affecter un sens à un mot, alors que l'IC se base sur des méthodes d'identification de concepts dans un corpus textuel (appariement de concept ou concept mapping).
- (2) la seconde raison est que, dans l'IC, la structure conceptuelle utilisée rend possible une extension de la représentation des documents (ou requêtes) via les différentes relations sémantiques qu'elle procure.

Nous donnons dans ce qui suit un résumé retraçant les apports des différents travaux concernant l'IS et l'IC en RI.

2.2.1.1. Indexation Sémantique (Sense Based Indexing)

L'IS est une approche d'indexation basée sur le sens des mots [Mihalcea et Moldovan, 00]. Elle repose sur des algorithmes de désambiguïsation de mots pour indexer les documents et les requêtes avec le sens des mots (mots-sens) plutôt qu'avec des mots simples. Une manière d'indexer serait par exemple, d'associer aux mots extraits, des mots du contexte qui aident à déterminer leur sens.

Des travaux cités dans [Baziz, 05] et [Boubekeur, 08] décrivent l'utilisation du sens de mots dans l'indexation, parmi ces travaux on trouve :

[Krovetz et al., 92] a pour but de trouver l'existence d'une relation entre, d'une part, la correspondance/non-correspondance des sens, d'autre part la pertinence/non-pertinence des documents restitués.

Dans [Voorhees, 93] le synset le plus approprié d'un mot ambigu est sélectionné à partir de WordNet par le calcul des nombres des mots communs entre le synset et les mots de contexte du mot à désambigüiser. Voorhees a indexé les documents et les requêtes par les synsets des noms pondérés par le schéma tf*idf.

Les travaux faites par [Mihalcea et Moldovan, 00] ont pour but d'identifier les termes d'indexations des documents dans l'ensemble des synsets de WordNet (termes synonymes définissant un sens d'un mot). Mihalcea et Moldovan, ont observé une amélioration de 16% dans le rappel et de 4% dans la précision quand ils ont utilisé une combinaison de l'indexation basée sur les mots clés et de l'indexation basée sur les synsets de WordNet. WordNet est utilisé pour déterminer l'appartenance des concepts d'une ontologie à un domaine donné, afin de garantir l'exploitation efficace des ontologies dans les moteurs de recherche [Hernandez et al., 08].

[Baziz et al, 04 ; 05] [Baziz, 05] proposent de représenter un document ou une requête par des concepts et des relations entre concepts. Cette approche consiste à projeter le contenu textuel d'un document (ou d'une requête) sur WordNet. L'objectif est d'extraire ses termes simples ou composés figurant dans WordNet. Baziz à proposer aussi une nouvelle formule de pondération « CF*IDF » (CF: Concept Frequency, IDF: Inverse of Document Frequency) adapter pour les termes composés.

Dans les démarches de [Xiaomeng et Atle, 06] et [Köhler et al., 06], le lexique de WordNet est utilisé afin de lemmatiser les termes.

[Boubekeur et al., 10b] proposent une approche d'IS similaire à [Boubekeur et al., 08] avec une différence au score de désambigüisation et au schéma de pondération des concepts. Le nouveau score de la désambigüisation d'un terme est calculé sur la base de fréquence de ce terme dans le document, et de ses distances sémantiques avec les concepts des autres termes les plus fréquents dans ce document. L'approche a été expérimentée sur la collection Muchmore[8], en utilisant le système Mercure qui est basé sur le modèle connexionniste [Boughanem et al, 92]. Les résultats rapportés présentent un gain de précision de plus de 50% avec les concepts pondérés par Tf*Idf, et des résultats moins précis avec le schéma de pondération des termes composés. [Boubekeur et al., 10b] expliquent cette diminution par le fait que le système Mercure est basé sur le schéma Tf*Idf.

[Harrathi et al., 10] proposent une approche d'IS de documents multilingues. Dans cette approche, Les termes simples sont extraits par la méthode d'indexation classique et les termes composés sont identifiés par une mesure statistique qui repose sur la fréquence des mots simples qui apparaissent mutuellement dans le contenu textuel d'un document ou d'une requête. Puis, un processus de désambigüisation ce déclenche pour les termes ambigus. L'approche de Harrathi est évaluée dans un SRI basé sur le modèle du langue proposé par [Maisonnasse et al., 09] et utilisant la ressource médicale UMLS[9] (Unified Medical Language System) et la collection de test CLEFmed2007[10] contenant des documents écrits en trois langues (anglais, français, allemand). Les résultats rapportés présentent un gain de précision moyenne de 5% par rapport à une indexation basée sur les mots clés.

[8] http:// muchmore.dfki.de/
[9] http://www.nlm.nih.gov/research/umls/
[10] http://www.clefcampaign.org/

[Mallak, 11] propose d'indexer les documents et les requêtes par des clusters de concepts les plus représentatifs de leurs contenus sémantiques. Il utilise la même technique de détection des concepts proposé par Baziz [Baziz et al., 05]. Pour la désambigüisation des concepts [Mallak, 11] a proposé une méthode basée sur la notion de centralité [Mallak, 11 ; Boughanem et al., 10].

Le système de [Azzoug et al., 11] commence par l'extraction des termes descriptifs (mots simples ou composés) à partir des documents (respectivement des requêtes) par le mapping de texte sur WordNet [Azzoug et al., 12]. Puis, la seconde étape consiste à trouver les sens correctes des termes ambigus déjà identifiés par une méthode de désambigüisation sémantique proposé dans [Azzoug et al., 13b] et basée sur WordNet et son extension aux domaines WordNetDomains [Magnini et al., 00]. La méthode de désambigüisation faite par rapport aux domaines d'usage en se basant sur l'idée que les mots de la langue, utilisés dans un même contexte, portent des sens fortement liés sémantiquement traitant un même domaine ou bien des domaines similaires afin de garder uniquement les sens liés au sujet du document par l'utilisation de WordNetDomains, puis utilise WordNet pour désambigüiser sémantiquement les termes de même domaine pour identifier le sens correct du mot dans son contexte, par attribution d'un score basé sur le cumul de ses similarités sémantiques dans WordNet avec les sens des autres mots de même domaine. Le sens approprié est le sens qui a le plus grand score. La dernière étape consiste à pondérer chaque concept par un poids traduisant son degré d'importance dans le texte, pour cela [Azzoug et al., 11] ont proposé deux schémas basés sur la notion de centralité d'un concept [Azzoug et al., 13a], la centralité d'un concept est traduite d'une part par son importance sémantique (exprimé par ses relations sémantiques avec les autres concepts du document) et d'autre part sa fréquence d'occurrence dans le document.

[Dinh, 12 ; Dinh et al., 10] présentent une approche d'IS pour le domaine biomédical en utilisant les concepts du thesaurus MeSh[11] (Medical Subject Headings). Cette approche commence par l'extraction des concepts à partir d'un document (respectivement une requête) en projetant son contenu textuel sur une liste préétablie de tous les concepts appartenant au thesaurus MeSh. Un score est ensuite affecté à chaque concept candidat du terme en se basant sur sa similarité thématique au texte et sa similarité structurelle définie par le degré de corrélation entre son entré dans le thésaurus et le contexte du terme dans le texte. La méthode s'occupe aussi de désambigüiser les concepts ambigus. Cette approche est évaluée sur la

[11] http://www.nlm.nih.gov/mesh/

collection des journaux médicaux OHSUMED[12]. Les résultats obtenus présentent un gain par rapport à ceux obtenus par indexation classique (un gain de performance de 17,35% pour la désambigüisation de proche en proche et de 17,06% avec la désambigüisation basée sur le clustering).

2.2.1.2. Indexation Conceptuelle

L'IC se base sur des concepts tirés d'ontologies et de taxonomies pour indexer les documents contrairement aux listes de mots simples.

Différents types d'ontologies sont utilisés dans le cadre de l'IC. Ces ontologies ne séparent pas les aspects de la connaissance liés au contenu des documents et ceux liés à la tâche de recherche réalisée [Hernandez et al., 08].

Ils existent des approches qui s'appuient sur des ontologies de domaine, ce qui permet de mieux spécifier le langage d'indexation. Parmi ces travaux nous trouvons :

Le projet Menelas vise à développer un système permettant d'accéder aux rapports médicaux de centres hospitaliers. Il repose donc sur une ontologie construite à partir des rapports à indexer qui modélise l'ensemble des maladies coronariennes [Zweigenbaum, 93].

Dans la hiérarchie de concept MESH (Medical Subject Heading), la mesure de similarité entre la représentation des requêtes et des documents donne l'avantage à l'ontologie dans le cas où elle est unique pour les deux représentations. La hiérarchie de concept MESH est utilisée pour indexer des documents de la médecine dans [Hearst et Karadi, 97].

khan et al., [khan et al., 04] Proposent une indexation basée sur des concepts d'ontologie de domaine du sport. L'approche commence par l'identification des termes d'indexation à partir du texte, puis la projection de ces termes sur les concepts de l'ontologie de domaine du sport pour déterminer les termes qui correspond à des concepts de l'ontologie. Pour désambigüiser les termes ambigus, khan et al., ont proposé une approche basée sur la distance sémantique par le calcul de score entre le concept ambigu et les autres mots de son contexte. Le score le plus élevé est retenu comme une distance minimale.

Le projet CADIS[13] cité dans [Kolar et al., 05] qui utilise le thésaurus multilingue EUROVOC (Hiérarchie de 8 niveaux de 6.000 classes touchant 21 domaines différents) comme source de vocabulaire a pour objectif la réalisation d'un outil d'aide à l'indexation

[12] http://trec.nist.gov/data/t9_filtering.html
[13] Computer Aided Document Indexing System

manuelle afin de représenter les documents d'un corpus d'une manière uniforme. CADIS n'effectue pas l'indexation automatique de documents, mais il rend plus facile la tâche à l'indexeur humain en fournissant des résultats des techniques de traitement statistique et de langage naturel intégrées.

Le système cité dans [Vallet et al., 07], représente les connaissances du domaine de la recherche sous forme d'ontologie. Cette ontologie, associée avec la gestion des préférences utilisateurs, permet d'enrichir les sémantiques évoquées au moment de la RI.

[Hernandez et al., 07] proposent un modèle de représentation des objets pédagogiques en utilisant trois ontologies : ontologie de thème pour leurs représentations sémantique, ontologie des tâches pour leurs usages dans les scénarios d'apprentissage et ontologie des théories pédagogiques.

[Chang et al., 07] utilise deux ontologies : Une ontologie noyau, construit à partir des métadonnées des ressources, qui représente la sémantique générale des ressources et une ontologie de domaine (Ex : Mathématique de la secondaire).

[Aufaure et al., 07] utilisent plusieurs ontologies complémentaires (une ontologie de domaine du tourisme construite manuellement et une ontologie de service) et WordNet. L'ontologie de service est reliée aux tâches du domaine et à chaque concept de l'ontologie de domaine correspond des services, tâches et activités.

Pour représenter le contenu des documents, [Boubekeur et al., 08] proposent une approche d'IC basée sur WordNet pour construire le graphe appelé CP-Net[14]. Après identification des termes (simples ou composés) d'indexation par l'approche classique, ils ont projeté ces termes sur WordNet afin de trouver toutes les entrées qui leurs correspondant. Ensuite, les termes simples sont pondérés par le schéma tf*idf et les termes composés par une mesure probabiliste des sens possibles de ces termes par rapport aux sens adjacents dans WordNet, en tenant compte de leurs fréquences d'occurrences dans le document. La désambigüisation des termes est calculée en fonction de la somme de ses similarités sémantiques avec les sens des autres mots dans le document, en tenant compte des poids de leurs termes respectifs. Finalement, le document (ainsi que la requête) est représenté par un graphe CP-Net, où les nœuds sont les concepts d'indexation retenus et les arcs représentent les relations

14. Graphe où les nœuds sont les concepts d'indexation et les arcs représentent les relations contextuelles latentes entre les concepts.

contextuelles latentes entre ces concepts trouvés en moyen des règles d'association sémantiques.

2.2.2. Appariement à partir d'ontologies

La phase d'appariement dans un SRI est une étape très importante pour juger la pertinence des documents regroupés pour une requête donnée. L'ontologie peut influencer dans ce processus afin de permettre au SRI de calculer la similarité (appelée aussi similarité sémantique) entre requête et document de façon approché.

L'appariement d'ontologie est souvent la recherche d'équivalence $A \equiv B$ entre deux concepts A et B de deux différentes hiérarchies. $A \equiv B$ si $A \subseteq B$ et $B \subseteq A$. La représentation des concepts comme une conjonction de concepts implique que les concepts ont la forme $B = B1 \cap ... \cap Bk$. Ainsi $A \subseteq B$ si et seulement si $A \subseteq Bi$, $\forall$ i=1 à k [Hernandez et al., 07].

La méthode présentée dans [Zhao et al., 07] utilise un algorithme de similarité qui prend en considération la représentation sémantique des requêtes et des titres de documents sous forme d'un arbre sémantique construit à partir d'une ontologie de domaine OWTS[15].

2.2.3. L'ontologie et la reformulation de la requête

Les utilisateurs d'un SRI ne maitrisent pas dans la plupart des cas le domaine recherché où ils expriment leurs besoins difficilement à l'aide d'une requête mal écrite. A ce point, l'ontologie intervient pour aider l'utilisateur à formuler sa requête par l'ajout de nouveaux termes et/ou ré-estimer leur poids afin d'exploiter efficacement la collection de document.

Il existe deux types d'expansion de requête dans la littérature : La première consiste à utiliser des ressources, internes ou externes, comme par exemple un dictionnaire [Moldovan et al., 99] ou bien WordNet [Voorhes, 94], pour l'extension des requêtes par l'ajout de nouveaux termes en relation avec les termes de la requête. La seconde solution est la réinjection de pertinence reposant sur l'analyse des termes contenus dans les documents jugés pertinents pour la requête initiale. L'idée est que l'ajout de termes liés aux termes initiaux de la requête peut permettre de retrouver des documents qui ne sont pas restitué auparavant.

[Harman, 92] a prouvé que la reformulation de requêtes a des effets positifs en RI. L'objectif de la reformulation est soit de limiter le silence soit de réduire les risques de bruit. Dans le premier cas, la requête est étendue à partir de termes similaires à ceux de la requête

[15] ontology-based weighted semantic tree similarity Algorithm

initiale. Dans le second cas la requête initiale est étendue ou modifiée à partir de termes qui ajoutent de l'information complémentaire à la représentation du besoin.

Dans (Ka)2 [Benjamins et al., 99], les pages Web sont annotées manuellement par des concepts d'une ontologie. Tous les concepts liés aux termes d'une requête donnée sont inférés et ajoutés à cette dernière. L'utilisateur est assisté dans la formulation ou le raffinement de sa requête à l'aide d'une interface proposée par ce système. L'utilisateur a la possibilité de naviguer dans l'ontologie et de centrer la visualisation sur la représentation des concepts qui l'intéresse comme il a été fait dans WebBrain[16].

Dans [Tomassen et al., 06], l'enrichissement de la requête utilisateur se fait par substitution des concepts de la requête par les vecteurs caractéristiques des concepts correspondants dans l'ontologie. Cette méthode associée à chaque concept de l'ontologie de domaine un vecteur caractéristique décrivant la similarité sémantique du concept avec les termes et concepts auxquels il est en relation (Synonyme, conjugaison, etc.) par rapport aux contenus des documents d'un corpus. Le but de ce système est de rapprocher les requêtes au contexte d'utilisateur et aux caractéristiques des collections de documents utilisant les ontologies.

[Aufaure et al., 07] adaptent le model vectoriel par la substitution des termes de la requête par des concepts de l'ontologie et classifie par service les résultats d'une requête en utilisant une ontologie de services permettant de spécifier les services liés à un domaine spécifique : acteurs, activités ou tâches. L'enrichissement de requêtes utilisateurs se fait par analyse morphologique et sémantique en utilisant les concepts et les relations entre l'ontologie de domaine et WordNet. L'utilisateur peut aussi utiliser l'ontologie de domaine pour désigner les concepts à utiliser dans sa requête.

Dans [Kim et al., 07], la RI se déroule en deux phases. Premièrement, la requête utilisateur est reformulée à l'aide des concepts de l'ontologie qui correspondent aux mots clés de la requête. Deuxièmement, le système réalise la recherche d'objet contenant ces concepts.

Un des difficultés fondamentales de la reformulation est la dimension de l'espace de recherche qui est élevé. La réduction de cet espace passe par la détermination des : [Efthimiadis, 96]

1. Critères de choix des termes d'extension,

[16] http://www.Webbrain.com/html/default_win.html_

2. Règles de calcul des poids des nouveaux termes,
3. Hypothèses de base quant aux liens entre termes et documents.

3. La désambigüisation des sens des mots

L'objectif principal de la WSD est de trouver le sens le plus correct des mots ambigus dans leurs contextes d'utilisation. De nombreuses approches de désambigüisation se trouvent dans littératures. Ces approches peuvent être divisées en deux catégories : les approches basées sur des ressources terminologiques comme des dictionnaires, thésaurus et les ontologies et les approches basées sur les corpus d'apprentissage qui se base sur des gros textes pour construire la connaissance nécessaire pour cela.

3.1. Les approches basées sur les ressources linguistiques

Ces approches se basent sur les dictionnaires informatisés, les thésaurus ou les ontologies pour désambiguïser un mot ambigu. Nous trouvons plusieurs travaux cités dans [Azzoug, 13] :

3.1.1. Les approches basées sur les dictionnaires informatisés

L'approche la plus connue basée sur les dictionnaires est celui de lesk [Lesk, 86]. Son principe consiste à identifier tous les sens possibles d'un mot ambigu à partir d'un dictionnaire. Puis, un score est attribué à chacun de ces sens par le calcul des mots communs entre la définition (gloss) de chaque sens du mot ambigu et les définitions des sens des mots de leur contexte. Le sens qui a le score le plus élevé est sélectionné comme un sens correct. Cette méthode peut rencontrer des problèmes si un mot ambigu possède le même nombre des mots communs dans leur définition.

Plusieurs chercheurs ont adoptés la méthode de Lesk dans leurs travaux, on trouve [Azzoug, 13]:

Wilks et al. [Wilks et al., 90] ont proposé d'étendre, le contexte et les sens d'un mot ambigu dans l'algorithme de lesk de façon manuelle par l'ajout des mots qui occurrent toujours avec les mots de contexte et les sens. Cette méthode est testée sur LDOCE[17] (Longman Dictionary of Contemporary English). Elle a donné un taux de performance égal à 45% par rapport à une désambigüisation manuelle.

[17] http://ldoce.longmandictionariesonline.com/main/Home.html

Une autre extension de l'algorithme de lesk à la base de réseau de neurones a été donnée par Véronis et al. [Véronis et al., 90]. Le but est de résoudre le problème de désambigüisation de plusieurs mots ambigus à la fois. Cette méthode utilise le dictionnaire anglais CED (Collins English Dictionnary). Cette méthode donne un taux de précision égal 71,7%.

[Guthrie et al., 91] adaptent l'approche de Wilks et al., à la différence que la définition d'un sens d'un mot caractérisé par une catégorie spécifique, est étendue seulement par l'ensemble des mots co-occurrents dans toutes les définitions assignées dans cette même catégorie dans le LDOCE.

[Cowie et al., 92] ajoutent un code de domaine attribué par LDOCE. Ce code est ajouté à la définition d'un sens, cette fois ci la désambigüisation d'un mot repose sur le nombre de mots et de codes communs. Cette approche donne un taux de précision égal à 47%.

3.1.2. Les approches basées sur un thésaurus

Les thésaurus sont caractérisés par la force de représenter les termes dans leur contexte. Le vocabulaire d'un thésaurus fournit une description sémantique des associations entre mots et classe les sens des mots liés sémantiquement dans des catégories sémantiques (catégories de domaines). Plusieurs travaux menés dans ce contexte tels que :

Les travaux de yarowsky [Yarowsky, 92] qui se basent sur les catégories sémantiques[18] du thésaurus Roget[19], pour désambigüiser les sens de l'encyclopédie Grolier multimédia. Cette désambigüisation consiste à déterminer la catégorie sémantique à partir du thésaurus par l'association des mots clés de la catégorie cible. Le sens adéquat est sélectionné à partir de la catégorie identifiée. Yarowsky a testé son approche sur 12 mots ambigus. Les résultats rapportés ont montré une précision de 92%.

[Mohammad et al., 06] ont proposé une approche basée sur le thésaurus Macquarie[20] pour la désambiguïsation des sens des mots. L'idée de base de leur approche est que la majorité des occurrences d'un mot dans un corpus textuel, ont le même sens, c'est le sens approprié. Le test de cette approche est fait sur un petit échantillon du corpus British National Corpus World Edition (BNC) [Burnard, 00]. Les résultats obtenus affichent un taux de précision supérieur à 50%.

[18] Le thésaurus Roget comporte 1024 catégories de domaines (telles que : ANIMAL/INSECT, TOOLS /MACHENERY, …etc) , qui couvrent les différents sens des mots.
[19] http://www.roget.org/
[20] http://www.macquariedictionary.com.au/anonymous@9c9B329512906/-/p/dict/index.html

3.1.3. Les approches basées sur une ontologie

Ces approches se basent sur les relations exploitées à travers les ontologies pour déterminer le sens le plus approprié d'un mot ambigu. Les approches de désambigüisation à travers les ontologies peuvent se diviser en deux classes : une basée sur l'ontologie linguistique tel que WordNet et l'autre sur l'ontologie de domaine.

Plusieurs travaux sont cités dans ce contexte :

[Sussna, 93] propose une méthode de désambigüisation des noms basée sur l'utilisation des relations de synonymie et antonymie de WordNet. Cette méthode est testée sur la collection TIME[21]. Elle a donné un taux de précision égal 56%.

[Banerjee et Pedersen, 03] adoptent la méthode de lesk pour prendre en charge les relations disponibles sur WordNet.

Un autre intérêt des ontologies est de permettre la désambiguïsation des termes de la requête. Dans [Guha et al., 03] la désambiguïsation se fait selon trois méthodes. La première consiste à choisir le concept dont les labels les plus fréquents dans les documents. La deuxième approche consiste à réaliser un profil utilisateur et à choisir le concept le plus proche de son profil. Finalement, la troisième prend en compte le contexte de la recherche et les documents recherchés par l'utilisateur.

[Köhler et al., 06] améliorent la désambigüisation des sens des mots en utilisant la lemmatisation des mots. De plus, ils proposent une méthode pour améliorer le rappel sans modifier la précision par l'utilisation des sous-concepts et super-concepts dans les différentes relations en respectant une certaine limite sur la profondeur des relations de subsomption.

Dans [Boubekeur et al., 10a ; 10b], le désambigüiseur se base sur la relation is-a pour désambigüiser les noms et les verbes. Ils ont proposé de désambiguïser un mot en s'appuyant sur des mesures de similarités sémantiques entres les synsets dans la taxonomie is-a des noms et verbes de WordNet.

Il existe d'autres travaux basés sur une ontologie de domaine. Ces travaux sont basés sur WordNet et son extension aux domaines WordNetDomains [Magnini et al., 00] :

[Gliozzo et al., 04] ont utilisé WordNet et WordNetDomains pour désambigüiser le mot ambigu par apport à son domaine. Ils ont comparé deux vecteurs, le premier contient des

[21] The Time collection consists of articles from the magazine Time.

synsets du mot ambigu extrait à partir de WordNet et le second, présente les domaines de ce mot dans leur contexte. Le sens adéquat est retenu par la similarité la plus élevée entre ces deux vecteurs. Cette approche a été évaluée sur la collection Senseval-221 et donne des résultats de précisions satisfaisant (79% pour la désambigüisation des verbes et noms) et (75% pour tous les catégories syntaxiques), et de faibles rappels (40% pour les verbes et noms et 35% pour les tous les mots).

Une méthode basée sur le même principe de la méthode précédente est proposée par Vázquez et al. [Vázquez et al., 04]. Ces chercheurs exploitent les domaines des mots de définitions (les glosses) dans le processus de désambigüisation. Cette méthode, testée sur la collection Senseval-2, produit un taux de précision de 47%.

[Kolte et al., 08 ;09] utilisent WordNet pour identifier le domaine du mot à désambigüiser dans son contexte. Kolte et al. ont testé cette approche sur SemCor et ils ont obtenu un taux de précession égal à 63,92%.

3.2. Les approches basées sur les corpus d'apprentissage

Ces approches se basent sur les techniques d'apprentissage. Elles sont divisées en deux catégories : approches supervisées et approches non supervisées.

3.2.1. Les approches supervisées

Ces approches nécessitent l'intervention de l'être humain pour annoter manuellement les textes de corpus d'apprentissage par les sens des mots. Plusieurs travaux existent dans ce contexte tels que :

[Weiss, 73] a utilisé un corpus étiqueté ADI[22]. Cette approche donne un taux de précision d'environ 90%.

Une autre approche similaire à celui de weiss est proposée par kelly [Kelly et al., 75] à la déférence que cette approche ne permet pas de désambigüiser une phrase complète.

L'approche de Yarowsky [Yarowsky, 00] est basée sur les arbres de décision pour identifier le sens adéquat du mot ambigu. Ce système de désambigüisation est considéré comme le meilleur système selon la compagne d'évaluation SENSEVAL de 1998, avec une précision de 78,9%.

[22] http://ir.dcs.gla.ac.uk/resources/test_collections/adi/

3.2.2. Les approches non supervisées

Ces approches se basent sur des corpus non annotés pour construire la connaissance nécessaire à la désambiguïsation. L'apprentissage est basé sur l'idée que les occurrences d'un mot qui ont un même sens possèdent souvent des mots co-occurrents similaires. Ces mots voisins sont regroupés en clusters. Ces clusters sont considérés comme des sens appropriés pour des mots ambigus. Parmi les travaux qui utilisent cette approche nous trouvons celle de Schütze [Schütze, 98] qui se base sur le modèle vectoriel.

4. Apport des ontologies dans les systèmes de recherche d'informations

De manière générale, ce qui est attendu d'une ontologie, est qu'elle assure la réutilisation de connaissances. En RI, son apport est ciblé. Nous donnerons dans ce qui suit d'après un rapport de [Masolo, 01], quelques résultats entendus de l'utilisation des ontologies dans les systèmes de RI [Baziz, 02]:

- **Les ontologies doivent réduire le silence dans les réponses aux requêtes :** le but est de trouver autant de documents pertinents que possible dans une collection donnée.
- **Les ontologies doivent aider à réduire le nombre de réponses bruitées**. L'idée est d'ignorer les documents contenant les mots de la requête, mais avec un sens différent.
- **Avec l'aide de l'ontologie, l'utilisateur peut exprimer son besoin plus facilement** : afin de guider l'utilisateur, des étapes peuvent lui être suggérées pour préparer sa requête ou une nouvelle formulation avec des termes plus appropriés.

5. Conclusion

Ce chapitre on a vu les critères utilisés pour choisir l'ontologie la plus adapté à un SRI. Nous avons décrit les différents systèmes existants et utilisant les ontologies : dans la reformulation de la requête, l'appariement ontologique, et la représentation des documents. Un mot dans un document (requête), est alors indexé différemment, selon le sens qu'il représente dans le contexte dans lequel il apparaît.

Dans ce type d'approche, le contexte d'un mot est souvent réduit à son voisinage immédiat (une fenêtre de quelques mots à gauche et/ou à droite du mot cible). Le mot-sens est alors représenté par le mot, auquel est associé, soit un numéro de sens tel qu'il apparaît dans une ressource sémantique externe, soit, d'autres mots de son contexte d'usage, permettant de le distinguer des autres sens.

Les autres travaux traitant de l'IC, quant à eux, s'emploient à attacher les termes des

documents (ou des requêtes) à des concepts de l'ontologie. Un des bénéfices que procurent ces approches, est d'exploiter l'opportunité de la présence de relations sémantiques entre concepts dans l'ontologie pour retrouver les documents pertinents. Nous avons décrit les différentes approches de désambigüisation qui sont basées sur les ressources linguistiques telles que (les dictionnaires informatisés, les thésaurus et les ontologies), ou bien les approches basées sur les corpus d'apprentissage (approche supervisé et non supervisé). Ces approches trouvent leur force dans le choix du sens le plus adéquat pour les termes ambigus. Dans ce qui suit, nous présenterons les travaux de recherche dans le cadre de l'utilisation des ressources sémantiques dans les systèmes de recherches d'informations arabes. Le but est de synthétiser tous les travaux existants dans ce domaine, afin de bien positionner nos travaux dans le domaine.

Table des matières

Chapitre 3 : La Recherche Sémantique pour les Textes Arabes : Etat de l'Art

1. Introduction

Par l'utilisation des ontologies, des thésaurus et des dictionnaires dans les SRI, la RI guidée par le sens a prouvé son efficacité pour la langue anglaise. La langue arabe a connu une augmentation très importante dans le volume des collections de documents, par conséquent, cela nécessite des SRI efficaces (plus intelligents) afin d'indexer ces collections et de trouver les informations pertinentes par rapport aux requêtes utilisateur.

La question qui se pose maintenant est la suivante : Est-ce que le changement de la langue pour un SRI implique une différence au niveau de la qualité du SRI ?

Pour répondre à cette question, nous allons étudier la problématique de l'utilisation des ontologies dans les SRI pour les textes arabes.

Ce chapitre est organisé comme suit : nous commençons par la description des caractéristiques et particularités de la langue arabe afin de dégager les problèmes spécifiques de la RI liés à la langue arabe. Nous abordons par la suite, les différents travaux concernant l'utilisation des ontologies dans les SRI arabes.

2. Les caractéristiques de la langue arabe

L'arabe, une des six langues officielles des Nations Unies, est la langue maternelle de plus de 300 millions de personnes [Dilekh, 11].

L'Arabe, langue sacrée du Coran, connaît une grande stabilité dans un créneau bien précis qui est celui de la littérature classique, des milieux de l'enseignement, la culture officielle et de la presse. C'est l'Arabe standard ou littéraire, universellement partagé par les lettrés de tous les pays arabes. Par contre, parallèlement à cette lignée, il existe de nombreuses branches s'écartant plus ou moins de la norme. L'Arabe dialectal dans toutes ses variétés, essentiellement oral, et le moyen Arabe (état intermédiaire entre le dialectal et le classique)

essentiellement écrit, sont autant de réalisations différentes d'une même source. Suffisamment proches pour constituer une seule et même langue, suffisamment éloignées pour ne pas s'intégrer dans les mêmes systèmes de traitement automatique [Zaidi, 13].

2.1. Particularité de la langue arabe

La langue arabe est composée de 28 lettres (voir tableau 3.1) (25 consonnes et 3 voyelles longues), les voyelles courtes n'étant pas représentées par des lettres mais par des diacritiques, placées sur ou sous les consonnes. Les lettres sont monocamérales, dans le sens où il n'existe pas de minuscule et de majuscule.

Lettre Arabe	Correspondant Français	Lettre arabe	Correspondant français
ا	A	ض	d
ب	B	ط	t
ت	T	ظ	z
ث	Th	ع	‘ ‘
ج	J	غ	gh
ح	H	ف	f
خ	Kh	ق	q
ﻠ	D	ك	k
ذ	D	ل	l
ر	R	م	M
ز	Z	ن	n
س	S	ه	h
ش	Sh	و	W
ص	S	ى	y

Tableau 3.1 : Les 28 lettres arabes

L'Arabe s'écrit de droite à gauche avec la particularité que les lettres épousent des formes différentes selon qu'elles soient au début, au milieu ou à la fin du mot, le tableau 3.2 illustre le script de quelques lettres dans les trois cas de graphie. Cependant, Il faut noter que certaines lettres ne s'attachent pas à celles qui la succèdent comme {ص، س، ر، د، ا}.

A la fin du mot	Au milieu du mot	Au début du mot
أ ، ؤ،ئ،ء	ـئـ	أ
ب، ـب	ـبـ	بـ
ـه، ه	ـهـ	هـ
ـم، م	ـمـ	مـ
ـي، ي	ـيـ	يـ
ـغ، غ	ـغـ	غـ

Tableau 3.2: Etat de transcription des lettres arabes

Des petits points noirs ont été utilisés comme marques de différenciation entre des lettres qui partageaient une forme identique. Ces points sont placés au-dessus et au-dessous de la lettre en un, deux ou trois. Exemples : ث , ت, ب; ق , ف; ج , خ , etc.

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

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

Les sept signes orthographiques sont :

- Sukun « ◌ْ » : ce signe indique qu'une consonne n'est pas suivie (ou muet) par une voyelle. Il est noté toujours au-dessus de la consonne ;

Les trois signes de tanwin : lorsque (la Fatha, la Kasra et la Damma) sont doublées, elles prennent un son nasal, comme si elles étaient suivies de « n » et on les prononce respectivement :

- an « ◌ً » pour les Fathatan ;
- in « ◌ٍ » pour les Kasratan ;
- un « ◌ٌ » pour les Dammatan.
- Chadda « ◌ّ » comme dans le français, l'arabe peut renforcer une consonne quelconque ;

- Wasla « ٱ » : quand la voyelle d'un Alif au commencement d'un mot doit être absorbée par la dernière voyelle du mot qui précède ;
- Madda « آ » : la madda (prolongation) se place sur l'Alif pour indiquer que cette lettre tient lieu de deux alifs consécutifs ou qu'elle ne doit pas porter le Hamza.

Cependant, les textes courants rencontrés dans les journaux et les livres ne comportent habituellement pas de voyelles. De plus, certaines lettres comme Alif « ا » peuvent symboliser le « أ », « إ », « آ » ou « ٱ »; de même que pour les lettres « ى » et « ه » qui symbolisent respectivement « ي » et « ة » [Dilekh, 11].

2.2. La structure morphologique d'un mot arabe

La définition du mot du point de vue du traitement automatique se heurte à des considérations syntaxiques et sémantiques. Dans le domaine des langages formels, la transformation du flux de caractères représentant un texte en une suite d'unités mieux adaptées aux traitements ultérieurs, est habituellement appelée segmentation (tokenization), et les unités produites les segments (tokens) sont construites sur la base de définitions purement orthographiques. En arabe cette séquence de lettres est appelée le mot graphique (MG). « Le mot graphique est facile à identifier : c'est ce qui s'écrit en un seul bloc entre deux blancs. » [Kouloughli, 94]

Un MG Arabe (MGA) peut être soit simple, soit complexe. Un MGA simple est un mot attesté de la langue, il est formé par la concaténation d'une base avec d'éventuels affixes (préfixes et suffixes). Il ne constitue pas un mot attestable de la langue sans les affixes [Abderrahim, 08].

MGA simple = Préfixes + Base + Suffixes

Un MGA complexe est formé par la concaténation d'un mot simple et un ensemble de clitiques (proclitiques et enclitiques).

MGA complexe = Proclitiques # mot simple # Enclitiques

MGA complexe = Proclitiques # Préfixes + Base + Suffixes # Enclitiques

Ou : MGA complexe = Prébases + Base + Postbases

Avec : Prébases=Proclitiques # Préfixes ; Postbases=Suffixes # Enclitiques

L'Arabe est une langue générative, les noms et les verbes sont dérivés d'une racine, généralement, trilitère. Nous pouvons engendrer jusqu'à 150 mots différents à l'aide de

schèmes et ce, à partir d'une même racine. Le tableau 3.3 donne quelques schèmes du mot « شهد ».

Schème	شهد	
فَعَلَ	شَهَدَ	Il a témoigné
فَعِلَ	شَهِدَ	Il a assisté
فَاعَلَ	شَاهَدَ	Il a regardé
فَاعِلٌ	شَاهِدٌ	Témoin
مَفْعَلٌ	مَشْهَدٌ	Scène
فُوعِلَ	شُوهِدَ	Il a été vu
فَعَالَةٌ	شَهَادَةٌ	Témoignage. Certificat
فَعِيل	شَهِيد	Martyr

Tableau 3.3 : Exemple des schèmes

Plusieurs types d'affixes sont agglutinés au début et à la fin des mots : antéfixes, préfixes, suffixes et post fixes (voir tableau 3.4).

Antéfixe	Préfixe	Noyau	Suffixe	Post fixe

Tableau 3.4 : Structure d'un mot

2.2.1. Les antéfixes

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

Voici une liste non exhaustive des antéfixes simples.

- La coordination par les coordonnants « فَ » fa et « وَ » wa.
- L'interrogation par le morphème «أَ » a.
- La marque du futur «سَ » sa.
- L'article «اَلْ » al.
- Les prépositions par les lettres « بِ » bi et « لِ » li.
- Les particules du subjonctifs النصب « فَ » fa, « لِ » li, et « وَ » wa.

- Le marqueur de comparaison par les lettres « كَ » ka.
- Le marqueur de corroboration « لَ » la.
- La particule du jussif (الجزم) par la lettre « لِ » li.

2.2.2. Les préfixes

Les préfixes (voir tableau 3.5), habituellement représentés par une seule lettre, indiquent la personne de conjugaison des verbes au présent.

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

Tableau 3.5: listes des préfixes arabes.

2.2.3. Les suffixes

Les suffixes sont les terminaisons de conjugaison des verbes et de marques duelles/plurielles/femelles pour les noms y compris les adverbaux. Ils ne se combinent pas entre eux. Voici la liste (tableau 3.6) exhaustive de tous les suffixes :

ا	ت	ك	و	يات
ات	ة	كم	وا	ية
اتكم	تان	ما	ون	يتنا
اتنا	تم	نا	ونن	يتها
اته	تموها	ني	ونه	ين
اتها	تنا	هـ	وه	يه
اتهم	ته	ها	وها	يها
اتية	تها	هم	وهم	يون
اها	تين	هن	يا	يين
ان	تهم	هما	ي	

Tableau 3.6: listes des suffixes arabes

2.2.4. Les post fixes

Finalement, les post fixes (voir tableau 3.7) représentent des pronoms attachés à la fin des mots. Ils peuvent se combiner entre eux. Voici dans le tableau suivant une liste des post fixes:

N° post fixe	post fixe	Description
1	ِي	1er Personne, Masculin/Féminin, Singulier
2	ي	1er Personne, Masculin/Féminin, Singulier
3	نَا	1er Personne, Masculin/Féminin, Duel/Pluriel
4	كَ	2eme Personne, Masculin, Singulier
5	كِ	2eme Personne, Féminin, Singulier
6	كَمَا	2eme Personne, Masculin/Féminin, Duel
7	كَمْ	2eme Personne, Masculin, Pluriel
8	كُنَّ	2eme Personne, Féminin, Pluriel
9	هُ	3eme Personne, Masculin, Singulier
10	هَا	3eme Personne, Féminin, Singulier
11	هُمَا	3eme Personne, Masculin/Féminin, Duel
12	هُمْ	3eme Personne, Masculin, Pluriel

13	هُنَّ	3eme Personne, Féminin, Pluriel
14	هِ	3eme Personne, Masculin, Singulier
15	هِماَ	3eme Personne, Masculin/Féminin, Duel
16	هِمْ	3eme Personne, Masculin, Pluriel
17	هِنَّ	3eme Personne, Féminin, Pluriel

Tableau 3.7: listes des post fixes arabes

Dans un mot arabe, la base est généralement entourée de propositions et de pronoms qui s'agglutinent à la racine en tant que préfixes, suffixes, infixes, antéfixes ou post fixes, de telle sorte qu'un mot arabe peut résumer à lui seul, toute une phrase exprimée dans une autre langue telle que le Français par exemple, le tableau 3.8 montre un exemple de segmentation d'un mot arabe.

أستمتلكونه : Est-ce que vous allez vous l'approprier ? Ce mot peut être segmenté ainsi :

ـه	ـونـ	لكـ	ـتـ	مـ	تـ	أسـ
			Infixe			
Postfixe	Suffixe	Corps schématique			Préfixe	Antéfixe
Pronom complément du nom	Suffixe verbal exprimant le pluriel	ملك: Racine			Préfixe verbal du temps de l'inaccompli	أ : Question سـ : Futur

Tableau 3.8: Exemple de segmentation d'un mot arabe

2.3. Les catégories du mot

Il existe trois catégories pour un mot arabe : nom, verbe et particule.

2.3.1. Le verbe

Le verbe est une entité qui exprime un sens variant en nombre, en personne et en temps, exemple : « كَتَبَ » ; sa conjugaison dépend du temps, du nombre, du genre, de la personne et du mode, il peut donc être exprimé à l'accompli ou l'inaccompli, au singulier, duel ou pluriel,

au masculin ou au féminin, au premier, deuxième ou troisième type et être au mode actif ou inactif.

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

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

La conjugaison des verbes dépend de plusieurs facteurs :

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

2.3.2. Le nom

Le nom est un élément désignant un être ou un objet qui exprime un sens indépendamment du temps, exemple : « مَكْتَبٌ ». Le nom peut être propre, commun ou dérivé d'un verbe. Il s'exprime au singulier, au duel ou au pluriel, au féminin ou au masculin. Il peut être agent, objet, instrument ou lieu.

Nous pouvons distinguer dans le tableau 3.9 deux classes de noms : la première regroupe les noms conjugables ou semi-conjugables qui peuvent avoir la forme duelle, plurielle, etc. la deuxième classe regroupe les noms non-conjugables qui gardent la même forme quelque soit le contexte. Les noms conjugables sont soit des noms primitifs, qui échappent à toute dérivation comme « غَزَالْ » (gazelle), soit des noms dérivationnels qui sont formés à partir d'une racine comme « مَكْتَبَةْ » bibliothèque) de la racine « كتب ».

Catégorie	Dérivation	Conjugaison	Sous-catégorie	Exemples
Nom	Dérivationnel irrégulier	Non conjugable	Adverbe	أين, حيث
			Nom de voix	كخ
			Nom de verbe	هيهات, أف, آه
			Pronom Personnel (affixé ou isolé)	هو, أنا, تن
			Pronom interrogatif	متى, ما
			Pronom conditionnel	من, إذا
			Pronom allusif	كَم, كأي
		Conjugable	Pronom relatif	الذي, التي
			Nom de nombre	خمسة, أربعة
			Pronom démonstratif	هذا, هذه
			Nom propre	محمد, عائشة
			Nom commun	رجل, قلم
	Dérivationnel régulier	Conjugable	Masdar	الحياة
			Participe actif	صائم
			Participe passif	وجود
			Nom d'une fois	جلسة
			Nom de manière	نظرة
			Nom de temps	مغرب
			Nom de lieu	مكتب
			Nom d'instrument	مسمار
			Adjectif	فحل
			Superlatif	أفضل
			Nom diminutif	كتيب
			Nom de relation	جزائري
			Intensif	غواص

Tableau 3.9: Classement des sous catégories de noms.

2.3.3. La particule

La particule est une entité qui sert à situer les événements par rapport au temps et par rapport à l'espace. Elles peuvent être des conjonctions de coordination « و, أو, أم... » ou de subordination « إذا, لأن.... ». Les particules sont généralement des mots outils, bien que jouant un rôle important dans la cohésion d'une phrase, sont souvent associées à des mots vides qui ne véhiculent pas un sens spécifique à un domaine donné. On distingue plusieurs types :

- Préposition : exemple (حتى, عن)
- Particules de coordination : exemple (ثم , أو, و)
- Particules interrogatives : exemple (ما,هل)
- Particules d'affirmation : exemple (نعم, بلى, أجل)
- Particules de négation : exemple (لا, لن, لم)
- Particules distinctive : exemple (أي)
- Particules relatives : exemple (ما)
- Particules de futur : exemple (سوف)
- Particules conditionnelles : exemple (إن, لو)

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

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

3. Les problèmes liés au traitement automatique de l'arabe

Vu ses particularités, le traitement automatique de l'Arabe, fait face à un certain nombre de problèmes, les plus importants sont le problème de la voyellation, l'agglutination et l'extraction de la racine.

3.1. Le problème de la voyellation

L'absence de la voyellation est très souvent une grande source d'ambiguïté pour l'analyse morphologique, syntaxique, sémantique et même pragmatique. La majorité des textes écrits, exception faite pour les textes sacrés et quelques ouvrages pédagogiques, sont non voyellés. Cette ambiguïté réside dans le fait que 74% des mots qui composent le vocabulaire arabe, acceptent plus d'une voyellation lexicale, et 89,9% des noms qui le constituent acceptent plus d'une voyellation casuelle. La proportion des mots ambigus passe de 90,5% si les comptages portent sur leurs voyellations globales [Debili et Achour, 98].

Si le problème est aussi commun au Français où 28 % des mots sont ambigus à cause de l'absence d'accentuation, en arabe la proportion est bien plus grande, en effet, l'ambiguïté touche 95% des mots [Douzidia, 04].

3.2. Le problème de l'agglutination

Une grande partie des mots arabes sont générés en agglutinant des proclitiques et des enclitiques à un radical. Pour déterminer un nom, par exemple, on ajoute (ا ل = al), comme dans le mot «الشمس» (Le soleil). Les pronoms personnels peuvent se rattacher aux noms (آياته = ses signes), comme aux verbes (أنزله = il l'a révélé). Les particules aux noms (كالمجرمين =sur le même pied d'égalité que les criminels), les conjonctions de coordination aux verbes (فتولى =et il se retira). Le problème, dans le cadre du traitement automatique de l'Arabe, est de pouvoir bien décomposer le mot en ses différentes parties [Baloul, 03].

3.3. L'extraction de la racine

Afin d'obtenir la racine d'un mot, il faut d'abord connaître le schème par lequel il a été dérivé, supprimer les éléments flexionnels (antéfixes, préfixes, suffixes, post fixes) qui lui sont attachés. En général des tables de préfixes et de suffixes sont utilisées. La nature agglutinative de l'Arabe rend cette tâche, assez difficile. Cette difficulté est encore plus accrue, lorsqu'il s'agit de textes non voyellés. L'analyse morphologique devra donc découper le mot et identifier des préfixes comme les conjonctions (و = et) et (ك = puis), des prépositions comme (ب = avec) et (لـ =pour), l'article défini (ال= le, la, les) et des suffixes de pronom possessif (ـه =à lui, ـها = à elle, ـهم =à eux, ـهن= à elles) etc. La phase d'analyse morphologique détermine un schème possible. Les préfixes et suffixes sont trouvés en enlevant progressivement des préfixes et des suffixes et en essayant de faire correspondre toutes les racines produites par un schème afin de retrouver la racine » [Douzidia, 04].

3.4. La terminologie

Le problème de terminologie dans la langue arabe cherche toujours sa solution. Il suffit de prendre comme exemple quelques termes linguistiques et informatiques improvisés sous plusieurs équivalents dans les différents pays arabes. Il est clair que ce problème engendre une autre difficulté dans le traitement automatique de l'Arabe [Zaidi, 13].

4. Problématique de la langue arabe et la recherche d'information

Les problèmes liés à l'indexation classique sont dus aux variations linguistiques de la langue utilisée, ces variations pour la langue arabe sont au nombre de trois :

- des variations morphologiques comme dans « مدرسة » et « مدرستان », « خيل » et « خيول » ;

- des variations lexicales (on utilise pour le même sens des mots différents) comme dans le cas dans « فرس » et « خيل » ;

Des variations sémantiques comme dans le cas de « الحجر : مرادف الصخر » et « الحجر : أنثى الخيل ».

L'utilisation des ontologies peut constituer une solution (parmi d'autres) pour résoudre le problème des variations sémantiques. Par ailleurs l'utilisation d'un analyseur morphologique peut suffire pour résoudre les deux premiers cas de variations morphologiques (صرفية) et lexicales (معجمية) [Abderrahim, 09a].

Dans les SRI classiques les variations lexicales réduisent la précision et augmentent le bruit [Egozi et al., 11].

Une solution à ces problèmes est d'utiliser l'aspect sémantique pour représenter les documents et les requêtes dans le processus de RI, suivant deux orientations : la première consiste à utiliser les mots complexes comme des unités d'informations [Hammache et al., 09]. La seconde comprend l'indexation sémantique et dans ce cas les concepts d'indexation sont construit par : (1) de la sémantique latente des textes ; (2) des sens associés aux mots par apport aux termes de contexte identifiés par les méthodes de désambigüisation (WSD) ; (3) ou bien par les concepts extraits à partir du texte (indexation conceptuelle).

5. La désambigüisation du sens des textes arabes

La langue arabe est connue comme une langue riche sémantiquement, un mot peut avoir plusieurs sens selon leur contexte d'utilisation, à cet effet, la désambigüisation devient une tâche importante afin de lever l'ambigüité des mots en question. Dans la littérature, nous trouvons peu de travaux de désambigüisation des textes arabes. Nous pouvons citer :

[Zouaghi et al., 11] ont proposé un changement dans l'algorithme de lesk par l'utilisation des mesures de similarités telles que; la mesure de Wu et Palmer basée sur la distance entre

deux noeuds de la hiérarchie et leur position par rapport à la racine, la mesure de Resnik, la mesure de Jiang et Conrath, la mesure de Lin et la mesure de Chodorow et Leacock; afin de trouver le gloss correspondant au sens du mot à désambigüiser pour les textes arabes. Cette technique utilise le dictionnaire « Al-Mujam al-Wasit » est testée avec le corpus « Latif-Al Sulaiti[23] ». D'après les résultats retournés la mesure de Leacock et Chodorow donne le meilleur taux de précision par apport à lesk d'origine, contrairement aux autres mesures qui ont donné des résultats minimales par apport à la méthode de lesk.

[Zouaghi et al., 12] ont proposé une méthode pour attribuer le sens exact pour un mot ambigu, cette méthode est basée sur la combinaison de méthodes de RI: mesures Harman, Okapi et Croft, avec algorithme de Lesk. Ce système doit résoudre l'ambiguïté sémantique lexicale dans la langue arabe.

Dans [Zouaghi et al., 12], les méthodes de RI sont utilisées pour estimer le sens le plus pertinent du mot ambigu. Cette estimation est basée sur le calcul de la proximité (sémantique score de cohérence) entre le contexte actuel (contexte d'apparition du mot ambigu), et les différents contextes d'utilisation de chaque sens du mot. La combinaison des mesures Harman, Croft et Okapi avec l'algorithme de lesk donne un taux de précision de 78%.

6. La Recherche d'Information pour la langue arabe

Dans cette partie nous allons décrire les travaux existant en recherche sémantique pour les textes arabes.

6.1. La langue arabe est l'indexation sémantique par des ontologies

Dans la littérature, il existe peu de travaux d'indexation sémantique pour la langue arabe, l'idée principale est d'exploiter une ressource sémantique (thésaurus, dictionnaire ou ontologie) afin de construite un index sémantique pour les collections des documents (respectivement requêtes), parmi ces travaux, nous pouvons citer :

[Tazzite et al., 08] ont utilisé des termes d'un dictionnaire terminologique arabe, construit à l'aide des experts il (couvre 20 domaines), pour indexer sémantiquement les collections des documents. L'indexation est faite par la liaison de tous les termes proche d'un domaine en se basant sur le dictionnaire produit par les experts. L'expérimentation est réalisée sur une

[23] http://www.comp.leeds.ac.uk/eric/latifa/research.htm

collection de 1000 documents, par l'utilisation de trois mesures de similarités Harman, Croft et okapi. Les résultats obtenus ont montré un peu d'amélioration en termes de précision et de rappel par apport au système classique.

L'approche développée dans [Gasmi, 09] consiste à créer une représentation conceptuelle pour les sites web en arabe. L'indexation est basée sur une ontologie de l'université (comme une ontologie de domaine) construite pour valider les concepts extraits à partir des pages web. L'approche de [Gasmi, 09] est certes bénéfique, néanmoins, elle n'a pas été évaluée en termes de précision et de rappel afin de quantifier l'apport réel de l'indexation sémantique dans les textes arabes.

[Bakhouche et al., 12] ont proposé de construire un espace conceptuel des documents basé sur les relations sémantiques (synonymie, homonymie,...) entre les mots, cet espace est représenté sous la forme d'un vecteur. Ce dernier représente l'index sémantique qui servira dans les applications comme la traduction ou le résumé automatique des textes arabes. Il faut noter toutefois, qu'aucune évaluation de ce modèle n'a été précisée dans ce travail.

[Achour et al., 13] ont utilisé trois ontologies pour faire l'indexation sémantique des ressources pédagogiques multilingues en ligne (arabe, français, anglais) : Une ontologie de domaine pour définir le domaine à apprendre/enseigner (informatique, électronique, physique,…), une autre pour décrire le thème à étudier (micro-électronique, électricité,…), par ailleurs, la troisième est utilisée pour décrire les différents types de ressources éducatives ainsi que leurs petites unités, qui peuvent être disponibles dans un environnement de e-Learning. L'objectif étant l'amélioration de l'accès et la recherche de ressources pédagogique.

6.2. La langue arabe est la reformulation des requêtes par des ontologies

Les travaux de Xu et al., [Xu et al.,02] montrent que l'utilisation d'un thésaurus améliore considérablement (18%) les performances d'un SRI arabe. Xu et al., ont montré aussi que l'utilisation d'une indexation basée sur les racines est plus performante que l'utilisation des schèmes pour les textes arabes.

Les travaux effectués dans Kanaan et al., [Kanaan et al.,05] ont montré que la reformulation manuelle par repondération des termes de la requête permet une amélioration des performances (rappel et précision) du SRI Arabe. Pour leur expérimentation, Kanaan et al., ont utilisé un corpus de 242 documents et un jeu de neuf (9) requêtes.

Les travaux de Hammo et al., [Hammo et al.,07] portant sur l'expansion de la requête par des termes issus d'un thésaurus montrent une amélioration dans le rappel du SRI arabe. Nous notons que le corpus utilisé était le coran.

Les travaux de Zaidi et Laskri [Zaidi et Laskri, 07] sur l'expansion de la requête en utilisant une ontologie du domaine juridique et WordNet ont permet d'obtenir des améliorations considérables dans les performances du SRI.

Le système de Ahmed et Nürnberger [Ahmed et Nürnberger, 08] propose d'assister l'utilisateur dans la reformulation de sa requête, par l'ajout des formes proches morphologiquement des formes de la requête initiale, en se basant sur le calcul de similarité des n-grams entre les mots de la requête initiale et ceux enregistrés dans un lexique. Etant basée sur la similarité des chaines de caractères, cette approche ne peut résoudre le problème des variations lexicales ou sémantiques. Pour les opérations d'indexation et de recherche, Ahmed et Nürnberger ont utilisé les services du moteur de recherche Google.

Le système de [Abderrahim, 09a] propose une interface d'expansion de la requête, cette interface utilise un analyseur morphologique afin d'extraire les racines des termes qui seront utilisées pour trouver les concepts similaires à partir d'une base lexicale (Wordnet arabe).

Les systèmes de [Abderrahim, 09b] et [Abderrahim, 09c] utilisent une ressource lexicale AWN et un analyseur morphologique pour reformuler (par extension) la requête de l'utilisateur. La requête élargie est envoyée à un moteur de recherche « Google ». Dans ce système, l'évaluation de la contribution réelle de l'enrichissement de la requête en arabe n'a pas été réalisée car elle est très complexe et nécessite par conséquent de lourdes expérimentations

Les travaux de Abderrahim et al., [Abderrahim et al.,10] qui se résument à l'utilisation d'une ressource lexicale (WordNet Arabe) et un analyseur morphologique pour reformuler, par expansion, la requête de l'utilisateur permettant d'améliorer le rappel, mais pas la précision du SRI.

L'étude faite par [Wedyan, 12] indique que les Questions/Réponses (QR) automatique basé sur un thésaurus peuvent améliorer la performance d'un SRI arabe de 10% à 20%.

[Abderrahim et al., 12] ont testé une technique de réinjection de la pertinence pour la reformulation de la requête dans un SRI Arabe. Leur expérimentation basée sur un corpus de

textes Arabe de taille moyenne. Les résultats obtenus montrent qu'il y a effectivement une amélioration globale dans les performances de SRI Arabe.

L'approche présentée dans [abderrahim, 13a] décrit dans le cadre de l'évaluation de la contribution réelle guidée par une ontologie lexicale dans le contexte d'un SRI arabe. L'expérimentation faite sur 2 corpus différents, par l'utilisation de 2 ressources linguistiques AWN et Arabic Dictionary of Meaning (ADM). L'objectif du processus d'enrichissement est la formulation d'une requête plus riche et plus précise. Ainsi, une conséquence directe est l'amélioration d'un SRI en renvoyant des résultats plus pertinents. Le système présente une amélioration de 4% lors de l'utilisation AWN par apport à ADM.

Le travail de [Abderrahim, 13b] est considéré comme une extension des travaux de [Abderrahim et al., 10] en proposant d'évaluer l'apport réel de la reformulation de la requête guidée par une ontologie dans un SRI Arabe. L'expérimentation faite présente une amélioration de 6% lors de l'utilisation AWN pour la reformulation des requêtes de l'utilisateur.

7. Synthèse

Les travaux présentés dans ce chapitre s'inscrivent dans le contexte général de l'utilisation de la sémantique pour la représentation de l'information dans les SRI pour les textes arabes et plus particulièrement dans le cadre de l'indexation sémantique des requêtes et des documents guidée par une ressource sémantique externe.

Une ressource sémantique externe est une structure pré-ordonnée qui peut se présenter sous forme d'ontologie, de hiérarchie de concepts, de réseau sémantique ou de thésaurus étendu. Le point commun à ses structures est qu'elles contiennent toutes un ensemble de concepts et de relations dénotant des liens sémantiques entre ces concepts. Le but est alors d'exploiter la sémantique contenue dans ces ressources, tout d'abord, pour une meilleure représentation de l'information et du besoin en information, puis, pour améliorer la correspondance entre le besoin de l'utilisateur et l'information.

Tous les travaux présentés dans ce chapitre encouragent les recherches dans le domaine d'utilisation des ontologies pour la représentation d'information pour la langue arabe, soit pour la reformulation du besoin d'information de l'utilisateur, soit pour l'indexation sémantique des collections de documents. Notre contribution s'inscrit dans le même axe de

recherche et elle vise à doter les SRI pour les textes arabes avec des outils permettant de prendre en charge la sémantique par l'utilisation des ontologies.

8. Conclusion

Dans ce chapitre avons décrit les caractéristiques de la langue arabe ainsi que les problèmes de la RI liés à cette langue. Nous avons terminé le chapitre par une présentation des déférents travaux utilisant les ontologies dans les SRI arabes.

Dans le chapitre suivant, nous présenterons notre contribution dans le cadre de cette thèse. Notre but étant de proposer une approche plus performante pour l'indexation des documents et requêtes dans un SRI pour les textes arabes.

Table des matières

Chapitre 4 : Recherche Sémantique pour les Textes Arabes : Contribution

Les travaux de ce chapitre sont publiés dans :

Abderrahim, Med-Alaeddine, A Med-El-Amine, and C Med-Amine. Using Arabic WordNet for Semantic Indexation in Information Retrieval System. International Journal of Computer Science Issues, 10(2) : p 327–332. (2013).

Mohammed Alaeddine Abderrahim, Mohammed Dib, Mohammed El-Amine Abderrahim & Mohammed Amine Chikh. Semantic indexing of Arabic texts for information retrieval system. International Journal of Speech Technology. ISSN 1381-2416, p 1-8. (2015).

1. Introduction

Dans ce chapitre, nous décrivons notre contribution en termes d'implémentation d'un ensemble d'outils permettant de prendre en charge la sémantique dans un SRI pour les textes arabes. Nous proposons aussi l'évaluation de l'approche proposée en terme de taux d'amélioration des performances d'un SRI pour les textes arabes.

2. Description de l'approche implémentée

Notre expérimentation est basée sur l'utilisation de WordNet Arabe pour indexer une collection des documents téléchargée du web et un ensemble de requêtes utilisateur construit manuellement. Un noyau de SRI (Lucene) existant sur le web a été étendu dans le but de permettre d'évaluer l'approche proposée.

2.1. Les ressources, corpus et outils utilisés

Dans cette section nous allons décrire :

- La ressource sémantique utilisée : WordNet Arabe.
- Le corpus de teste pour l'évaluation de l'approche proposée : il comprend une collection de documents, un ensemble de requêtes construit manuellement avec l'ensemble des réponses supposées répondre à ces requêtes.

- Un ensemble d'outils réalisé, permettant de prendre en compte la partie IS du SRI et la partie évaluation de l'approche proposée.

2.1.1. WordNet Arabe

Nous avons utilisé WordNet Arabe qui est une base de données lexicale librement disponible pour l'arabe standard. Cette base de données suit la conception et la méthodologie du Princeton WordNet pour l'anglais et d'Euro-WordNet pour les langues européennes. Sa structure est celle d'un thésaurus, il est organisé autour de la structure des synsets, c'est-à-dire des ensembles de synonymes et de pointeurs décrivant des relations vers d'autres synsets. Chaque mot peut appartenir à un ou plusieurs synsets, et à une ou plusieurs catégories du discours. Ces catégories sont au nombre de quatre : nom, verbe, adjectif et adverbe. WordNet arabe est donc un réseau lexical dont les synsets sont les nœuds et les relations entre synsets sont les arcs. Il faut noter toutefois que WordNet Arabe est une des rares ressources « libre » pour la langue générale arabe disponible en ligne. Actuellement (septembre 2015), WordNet Arabe est dans sa version « 2.0 », Il compte 11269 synsets (7960 noms, 2538 verbes, 661 adjectifs et 110 adverbes), et 23481 mots [Elkateb et al., 2006a,b ; Black et al., 2006 ; Abouenour et al., 2013].

WordNet arabe est librement téléchargeable sur internet sous la forme d'une base de données relationnelle avec une interface d'accès en Java. Cette version est nommée AWNBrowser_2.0.1[24] (voir interface d'accès de WordNet arabe figure 4.1).

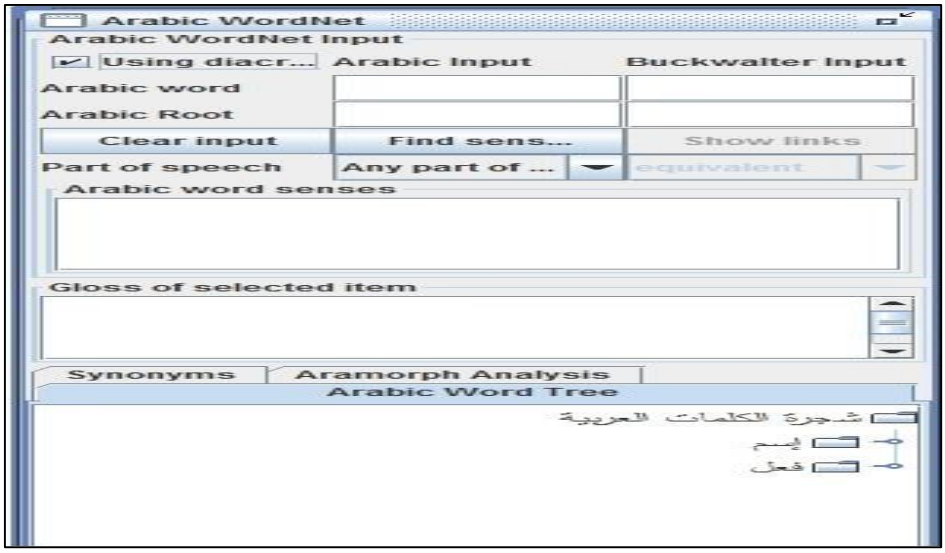


Figure 4.1 : Interface de WordNet Arabe

[24]http://sourceforge.net/projects/awnbrowser/

2.1.2. Corpus d'évaluation

Pour notre expérimentation nous avons utilisé une collection de plus de 22 429 documents arabe (environ 180 Mo) de différents domaines (voir tableau 4.1). Cette collection compte environ 17 000 000 mots dont 612 650 mots différents. Cette collection est divisée en plusieurs domaines comme suit :

Les domaines du corpus	La taille du corpus par domaine (M.O)	Le Nombre des documents par domaine
Astronomie	3,25	557
Droit	8.2	944
Economie	19,3	3102
Education	35,8	3608
Histoire	46,3	3233
Histoires	12,9	726
Recette	4,0	2373
Religion	19,8	3171
Santé	18,6	2296
Sport	9,75	2419

Tableau 4.1 : La collection des documents

Les textes de cette collection sont organisés sous la forme de fichier textuel écrit en arabe (voir un exemple de fichier texte dans la figure 4.2) :

183وسيلة دعوية للمرأة المسلمة ... سنبلة الحفيدة183 وسيلة دعوية للمرأة المسلمة سنبلة الحفيدة * مدارس كثيرة بها خيرة المعلمات علما ودعوة ونشاطاً.. أما هي عندما عينت فإنها ارسلت إلى مدرسة نزخر بالمعلمات لكنهن نائمات.. فلا توجد محاضرات ولا دروس.. في البداية بدأت في التودد إلى المدرسات وقالت: هم أهم عندي الآن من الطالبات لأنهن داعيات خاملات آتزن الكسل والدعة.. فقط بحتجن إلى إيقاظ.. بدأت خطوات الإيقاظ بالكتاب والشريط والهدية.. حتى تحولت المدرسة إلى خلية نشاط ومركز دعوة.. حمدت إحداهن الله وهي تردد كيف ضاعت مني خمس سنوات يومياً أقف أمام الطالبات ولم أدعهن وأحدثهن وأركز على تربيتهن!! إنها الخطة اليوم والحساب غداً. * شرعت المعلمة في بيان أضرار السفر والفساد والانحلال في بلاد الكفر.. وعقبت بدعاء صادق.. نسأل الله أن لا ندخلها ولا نذهب إليها.. ولم ينته الدعاء حتى تسلل من بين الصفوف صوت حمل هم الدعوة: نعم يا معلمة.. نسأل الله عز وجل ألا ندخلها إلا فاتحين!! لا فض الله فوك وجعلك وأبناءك من الفاتحين. * نكد ونكدح للآخرة...- والله- إنها تركض للآخرة ركضا ونسعى الا سعياً.. فمن محاضرات إلى ندوات إلى نصائح. كل عمل خير لها فيه نصيب. وفي نهاية كل شهر- علمت إحدى المدرسات من زميلاتها- أنها ترسل راتبها كاملاً لأعمال الخير.. نعم كاملاً وأقسمت لقد رأته- برباطه- ترسل به إلى كفالة أيتام وطبع كتب وتجهيز غاز.. جعل الله مستقرك جنات عدن أيتها المؤمنة ورفع درجتك وأعلى منزلتك وكثر من أمثالك.. ووالله لأنت حفيدة عائشة وفاطمة. * اجتمعت معلمات المدرسة وقررن الدخول في (جمعية) مع بعض.. وكل منهن تتحدث عما ستفعل بالمبلغ عندما تستلمه أما هي فصامتة تنتظر ذلك اليوم.. حتى إذا استلمت المبلغ دفعت به لبناء مسجد.. لعله بصيبها الأجر والثواب

Figure 4.2 : Exemple d'un fichier de texte de la collection

Il faut noter que ces textes ne sont pas voyellés.

Nous avons utilisé un ensemble de 70 requêtes simples de différents domaines que nous avons construit manuellement (voir figure 4.3 pour des exemples de requêtes). L'ensemble des documents pertinents de la collection pour chaque requête construite a été enregistré dans un fichier à part. Ce dernier va nous permettre de faire les calculs de précision et de rappel.

N°	Requête
1	تربية
2	تسامح
3	تصوير
4	تضخم

Figure 4.3 : Exemples de requêtes utilisateur

2.1.3. Lucene

Lucene[25]est un moteur de recherche API libre, c'est une bibliothèque qui intègre toutes les fonctionnalités d'un vrai moteur de recherche classique (indexation et recherche par mot clé). Lucene est écrit principalement en Java et pour lequel il existe de nombreux portages dans d'autres langages de programmation comme C, C++, Perl, etc. Lucene supporte la langue arabe dans ses nouvelles versions (3.5.0 et plus) avec un analyseur propre à cette langue. La construction de l'index passe donc par cet analyseur et la consultation de WordNet arabe pour extraire les concepts proches. Le module d'appariement de Lucene demande lui aussi une extension pour l'utilisation de WordNet arabe.

Nous avons utilisé Lucene dans notre travail soit pour : (1) Effectuer une indexation simple sans utilisation WordNet arabe. C'est une indexation à base des mots clés. Nous avons implémenté ce traitement dans le but de faire la comparaison des résultats entre un SRI à base des mots clés et le SRI de notre approche ; (2) Effectuer une IS en utilisant AWN.

2.2. Les traitements proposés

Dans cette partie nous allons décrire les algorithmes proposés et implémentés pour la désambiguïsation et l'IS des textes dans le cadre d'un SRI pour les textes arabes.

[25] http://www.apache.org/dyn/closer.cgi/lucene/java/

2.2.1. La désambiguïsation

Nous avons étudié et implémenté deux algorithmes pour la désambiguïsation :

- L'algorithme de désambiguïsation par le concept commun.
- L'algorithme de désambiguïsation de Lesk.

2.2.1.1. La désambigüisation par le concept commun

Cette proposition est publiée dans le [Abderrahim et al., 2013]. Dans ce cadre, l'algorithme de désambigüisation cherche le concept le plus reliés aux concepts de même document, c-a-d, si un terme figure dans deux synsets ou plus nous choisissons le concept commun entre ces synsets s'il existe, sinon, l'algorithme choisi le synset qui regroupe le plus de concepts trouvés dans le document. Le déroulement de cet algorithme, schématisé dans la figure 4.4, est expliqué par l'exemple (1).

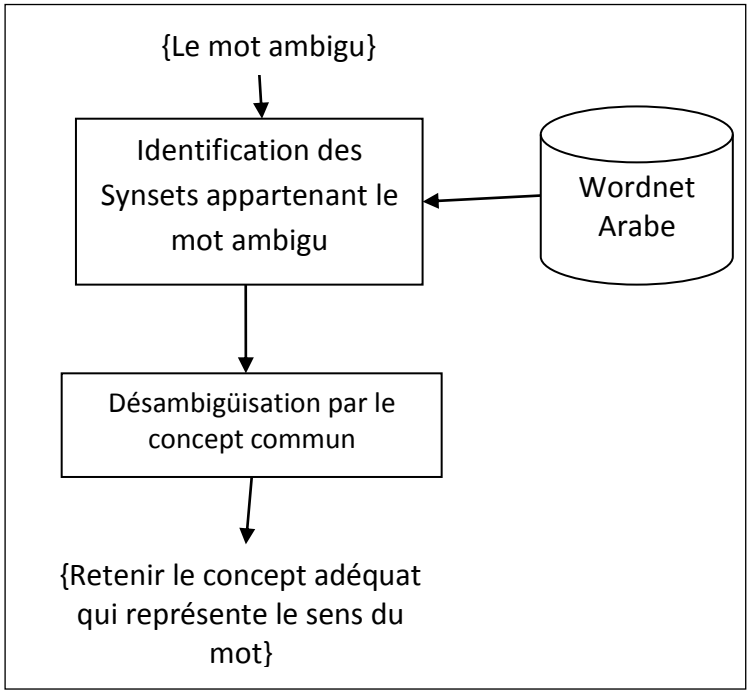


Figure 4.4 : Désambigüisation par le concept commun

L'algorithme de désambigüisation par le concept commun se présente ci-dessous comme suit :

Algorithme. Désambigüisation par le concept commun

1: **Entrée :** M_A *// Mot graphique à désambigüiser*

2 : *$S \leftarrow \{S_1, \ldots, S_N\}$ // les synsets candidats du mot à désambigüiser*
M_Texte // La liste des mots du texte à indexer

3: **Sortie :** *best_candidate // le sens le plus approprié pour le mot à désambigüiser*

4 : **Début**

5 : **Pour tout** *sens à désambiguïser M_A* **faire**

6 : *best_score ← 0 // c'est la valeur attribué au meilleur sens*

7 : *best_candidate ← Ø // initialiser à l'ensemble vide*

8 : *Sup ← 0 // le nombre de superpositions entre les Synsets du*
// mot à désambigüiser et les mots du texte à désambigüiser.

9 : **Pour tout** *j de 1 à N* **faire**

10 : *best_candidate ← $S_{(j)} \cap S_{(j+1)}$ // Calculer le chevauchement entre chaque définition*
// possible de le mot ambigu et les définitions des mots contenus dans leur contexte.

11 : **Fin boucle**

12 : **Si** *best_candidate = Ø* **Alors**

13 : **Début**

14 : **Pour tout** *j de 1 à N* **faire**

15 : *Sup ← $|S_{(j)} \cap M\_Texte|$ // Calculer le chevauchement entre chaque Synset*
// possible de le mot ambigu et les texte à désambigüiser.

16 : **Si** *Sup > best_score* **Alors**

17 : **Début**

18 : *best_score ← sup // Attribuer la meilleur valeur pour le j éme synset*

19 : *best_candidate ← $S_{(j)} \cap M\_Texte$ // Remplacer la valeur de best_candidate par le*
//nouveau sens

20 : **Fin condition**

21 : **Fin boucle**

22 : **Fin condition**

23 : **Fin boucle**

24 : **Fin Algorithme**

- ✓ **Exemple de désambigüisation par le concept commun**

Dans ce qui suit nous donnons un exemple décrivant le principe de la méthode de désambigüisation utilisée : En entrée nous avons le texte écrit en arabe suivant :

" سواء كانت حالة فقدان الذاكرة بشكل مؤقت أو دائم، أو جاءت بشكل مفاجئ أو ببطء فذلك يعتمد علي أسباب حدوث فقدان الذاكرة. إن عملية تقدم العمر قد ينتج عنها صعوبة في تعلم أو إدراك الأشياء الحديثة علي الشخص أو يمكن أن تتسبب في استغراق وقت أطول من قبل الشخص المسن في تذكر أو استدعاء الأشياء الحديثة عليه (ولكن التقدم في العمر لا يكون سبب في فقدان الذاكرة إلا إذا كان هذا التقدم مصحوباً بمرض معين ساعد في حدوث هذه الحالة). "

En sortie l'algorithme produit les concepts (index choisi) sélectionnés du tableau ci-dessous (voir colonne index choisi du tableau 4.2)

Termes	**Synsets sélectionnés à partir AWN**	**Entrée de Synset (Index choisi)**
حدوث	{ وُقُوع, ظُهُور, حُدُوث, حُصُول, حَدَث}	حُصُول
	{ وَاْقِع, حَدَث, حَادِثَة, حُصُول, حُدُوث}	
استدعاء	{ تَذَكُّر, اِسْتِدْعاء, ذِكْرَى}	تَذَكُّر
	{ طَلَب حُضُور, اِسْتِدْعاء}	
تذكر	{ تَذَكُّر, ذَاكِرَة}	ذَاكِرَة
	{ تَذَكُّر, اِسْتِدْعاء, ذِكْرَى}	
جاء	{ جَاءَ,أَتَى}	أَتَى
	{ ظَهَرَ,جَاءَ}	
	{ أتى,حَضَرَ,جَاءَ,قَدِم}	
ذاكرة	{ فِكْر, ذَاكِرَة }	تَذَكُّر
	{ ذَاكِرَة,تَذَكُّر}	

Tableau 4.2 : Exemple de sélection des concepts à partir de AWN par la méthode du concept commun.

2.2.1.2. La désambiguïsation par Lesk

Cette proposition est publiée dans le [Abderrahim et al., 2015]. En effet, l'algorithme de Lesk permet de faire la désambigüisation du sens en s'appuyant sur un principe simple. Cet algorithme exploite la notion de similarité entre les définitions reliées au terme à désambigüiser et les définitions reliées aux termes de leur contexte. Le résultat de cet algorithme est le sens le plus approprié au terme à désambigüiser. Cet algorithme a été intégré dans notre processus d'IS afin de La figure ci-dessous (voir figure 4.5) schématise l'algorithme de Lesk [Vasilescu, 2003].

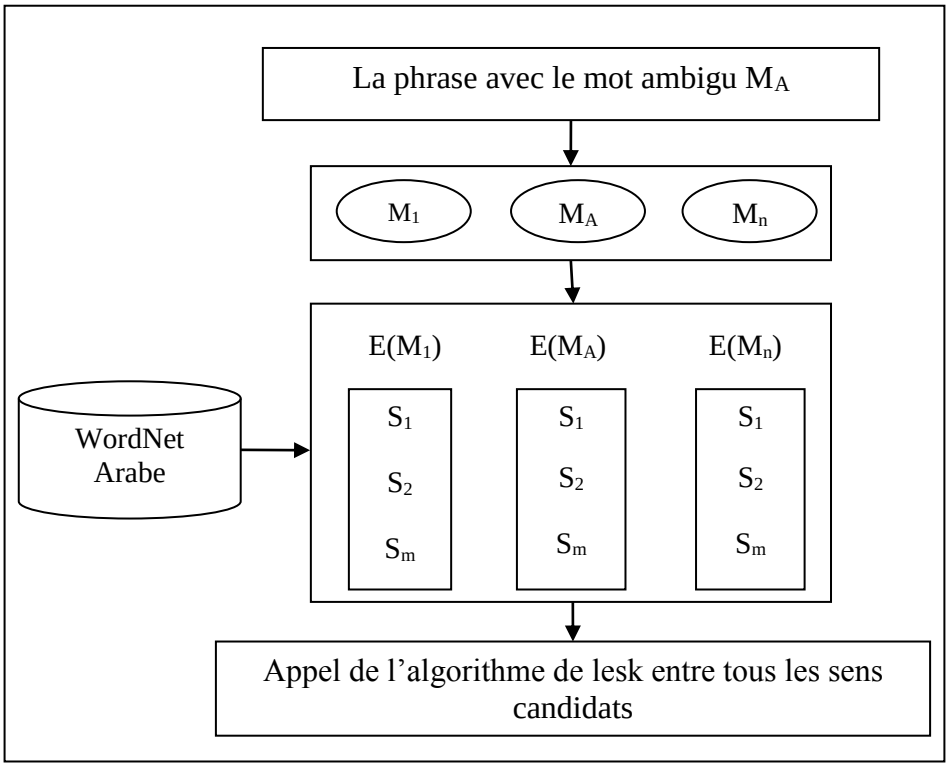


Figure 4.5 : Schéma descriptive de l'algorithme de Lesk

Où :
M_i : représente mot graphique[26] « i » à désambigüiser,
S_i : représente un sens candidat du mot graphique « i » à désambigüiser,
$D(S_i)$: représente la définition du sens candidat du mot graphique « i » à désambigüiser,
$E(M_j)$: représente la définition du sens du mot graphique « j » du contexte du mot graphique « i » à désambigüiser,

L'algorithme de Lesk se présente ci-dessous comme suit :

Algorithme. Désambiguïsation de Lesk

1: **Entrée :** *M_A // Mot graphique à désambigüiser*

2 : *$S \leftarrow$ {S_1,…, S_N} // les sens candidats du mot à désambigüiser ordonnés par // ordre décroissant de leurs fréquences*

3: **Sortie :** *best_candidate // le sens le plus approprié pour le mot à désambigüiser*

4 : **Début**

5 : **Pour tout** *sens à désambiguïser M_A* **faire**

[26] Par mot graphique on entend toute suite de caractères séparée par deux blancs.

6 : *best_score* ← *0* *// c'est la valeur attribué au meilleur sens*
7 : *best_candidate* ← S_1 *// initialiser avec le sens le plus fréquent à partir de la liste*
8 : *Sup* ← *0* *// le nombre de superpositions entre les définitions des sens du*
// mot à désambigüiser et les définitions des mots de leur
// contexte
9 : $C(M_A)$ ← $\{M_1, M_2, \ldots, M_i, M_N\}$ *// $C(M_A)$ représente le contexte de M_A : consiste à*
// prendre les mots adjacentes du mot à
// désambigüiser (M_A) à partir du texte
10 : **Pour tout** *j de 1 à N* **faire**
11 : $D(S_j)$ ← *Définition*(S_j) *// Extraire du AWN la définition de chaque sens candidate*
// S_j
12 : $E(M_j)$ ← *Définition*(M_j) *// Extraire du AWN la définition de chaque mot du contexte*
// M_j
13 : *Sup* ← $|E(M_j) \cap D(S_j)|$ *// Calculer le chevauchement entre chaque définition*
// possible de le mot ambigu et les définitions des mots
// contenus dans leur contexte.
14 : **Si** *best_score* < *Sup* **Alors**
15 : **Début**
16 : best_score ← *sup* *// Attribuer la meilleur valeur pour le j éme sens*
17 : best_candidate ← S_j *// Remplacer la valeur de best_candidate par le nouveau sens*
18 : Fin condition

19 : Fin boucle

19 : Fin boucle

20 : Fin Algorithme

- ✓ **Exemple de désambiguïsation avec l'algorithme de Lesk**

Dans ce qui suit nous donnons un exemple de désambiguïsation de Lesk :

En entrée nous avons le texte écrit en arabe suivant :

"غالبا ما يلجأ الطبيب المختص إلى العملية الجراحية لعلاج المريض"

Le tableau 4.3 présente les mots graphiques, les synsets correspondants ainsi que le sens choisi après l'opération de désambiguïsation avec l'algorithme de Lesk du texte en entrée. Il

faut noter que les mots graphiques de départ ont subi une opération de lemmatisation afin de ne retenir que les lemmes. Nous remarquons aussi que les mots graphiques {إلى ,ما} ne figurent pas dans le tableau puisqu'ils représentent des mots outils.

Mot graphique	**Synsets correspond à partir de WordNet**	**Sens choisi après désambiguïsation avec l'algorithme de Lesk**
غالبا	{غالبا,بشكل إعتيادي,بشكل عادي, عادة }	عادة
يلجأ	{ آوى,ألجأ }	آوى
طبيب	{طبيب} {طبيب, طبيب ممارس}	طبيب
مُخْتَصّ	{اخصائي, اختصاصي, متخصص, مختص}	اخصائي
عملية	{عملية} {عمل,عمل عسكري,عملية,عملية عسكرية,معركة} {عملية,عملية عسكرية} {إجراء, عملية, طريقة} {عملية,عملية جراحية,جراحة} {عملية,عملية معرفية,عملية إدراكية} {عملية, عملية لا شعورية, عملية غير مقصودة}	جراحة
جراحية	{عملية,عملية جراحية,جراحة}	جراحة
علاج	{علاج,معالجة,مداواة} {علاج,دواء,معالجة} {علاج,دواء}	مداواة
مريض	{شخص مريض,مريض,مصاب,متوعك}	شخص مريض

Tableau 4.3 : Exemple de sélection de concepts à partir de AWN par Lesk

Pour la méthode de désambigüisation nous avons choisi un contexte égal à 6 car c'est celui qui donne des meilleurs résultats d'après [Zouaghi et al., 2012]. Par exemple pour le mot graphique « عملية» et son contexte «المريض لعلاج الجراحية العملية إلى المختص الطبيب يلجأ». Après les étapes de prétraitements (élimination des mots vides {... ,إلى ,ما}, suppression des suffixes et préfixes, nous appliquons l'algorithme de Lesk. Nous calculons le nombre des définitions communs entre les Synsets correspondant aux termes restants « يلجأ, طبيب, مختص, عملية, جراحية, علاج, مريض », voir le tableau (4.4) :

Les termes du contexte							**Nombre des définitions communs**	**Sens choisi**
يلجأ	**طبيب**	**مختص**	**عملية**	**جراحية**	**علاج**	**مريض**		
أوى,ألجأ	طبيب	اخصائي, اختصاصي, متخصص, مختص	عملية	عملية,عملية جراحية, جراحة	علاج ,معالجة, مداواة	شخص مريض, مريض, مصاب, متوعك	1	/
			عمل,عمل عسكري,عملية,عملية عسكرية,معركة				1	/
			عملية,عملية عسكرية				1	/
			إجراء, عملية, طريقة		علاج ,دواء, معالجة		1	/
	طبيب, طبيب ممارس		عملية,عملية جراحية,جراحة				3	جراحة
			عملية,عملية معرفية,عملية إدراكية				1	/
			عملية, عملية لا شعورية, عملية غير مقصودة		علاج, دواء		1	/

Tableau 4.4 : Déroulement de L'algorithme de Lesk

2.2.2. L'indexation sémantique

L'algorithme de l'IS consiste à parcourir la liste de tous les mots graphiques appartenant à chaque document de la collection des documents (respectivement requête) et à vérifier s'ils

peuvent représenter des concepts qui appartiennent à WordNet arabe. Le choix des concepts représentatifs impose un ensemble de traitements afin de valider le concept ajouté à l'index final. Ces traitements se devisent en deux parties ; (1) un ensemble de prétraitements pour préparer les termes du texte à la seconde étape. (2) un ensemble de traitements pour l'identification des concepts dans Wordnet arabe. Puisque le processus d'IS des documents est identique aux requêtes, alors la figure 4.6 ne représente que celui des documents.

Les prétraitements ont pour but d'extraire un terme qui peut être présent dans Wordnet arabe. Ces traitements commencent par la segmentation du texte en un ensemble d'unités appelé mots graphiques, puis le système supprime les mots vides en comparant la liste des mots graphique à une liste prédéfinis des mots inutiles.

A la fin du processus précédent, nous obtenons un ensemble de termes qui serviront comme entrée au processus d'identification des concepts. Ce dernier prend chaque terme de la liste et identifie le synset lui correspondant dans Wordnet arabe. Ce traitement est complété par d'autres traitements ; dans le cas où un terme se situe dans deux synsets différents (ambigüité), dans ce cas le processus d'identification a besoin d'une désambigüisation des termes ambigus. La désambigüisation est réalisée en utilisant les deux méthodes abordées dans le début de ce chapitre. Dans le cas où le terme (dans sa forme) en question n'existe pas dans Wordnet arabe, alors un traitement supplémentaire est nécessaire pour extraire la forme de base pour le terme.

Les concepts trouvés forment l'index sémantique. Pour compléter l'information exprimée par l'index généré à partir des documents, nous ajoutons les termes qui n'appartiennent pas à WordNet arabe. Le processus d'indexation se déroule selon l'algorithme suivant :

Algorithme. IS des documents

1: **Entrée :** *D //Collection des documents*

2 : AWN // WordNet Arabe

3 : Lv // liste finie des mots outils {…. ,أينما ,ما ,و}

4: **Sortie :** *Index // Structure représentant l'index de la collection des documents*

5 : **Début**

6: **Pour tout** *di ∈D* **faire** *// di←{m1, m2,…., mi }*

7: **Pour tout** *mi ∈ di* **faire** *// mi : Mot graphique, une suite de caractère séparée*
// par deux blancs

8 : **Si** *mi ∈Lv* **alors** *Supprimer (mi) // On supprime tous les mots outils (vides)*

9 : **Sinon** *// mi ∉Lv*

10 : **Début**

11 : *Sys (mi) ← ∅ // Ensemble de synsets de mi*

12 : **Si** *mi ∉ AWN* **Alors** *mi ← Racine (mi) // extraire la racine de mi*

13 : **Si** *mi ∈ AWN* **Alors**

14 : **Début**

15 : *Sys (mi)← ensemble des sysnsets de AWN*

16 : **Si** *Card(Sys (mi)) > 1* **Alors** *// Card(…) représente la cardinalité*

17 : **Début**

18 : *Sys (mi) ← Désambigüiser (Sys (mi))*

19 : *Index← Index ∪Sys (mi)*

20 : **Fin condition**

21 : **Sinon Si** *Card(Sys (mi)) == 1* **Alors**

22 : *Index← Index ∪Sys (mi)*

23 : **Fin condition**

24 : **Sinon** *Index← Index ∪mi // mi ∉ AWN*

25 : **Fin condition**

26 : **Fin boucle**

27 : **Fin boucle**

28 : **Fin Algorithme**

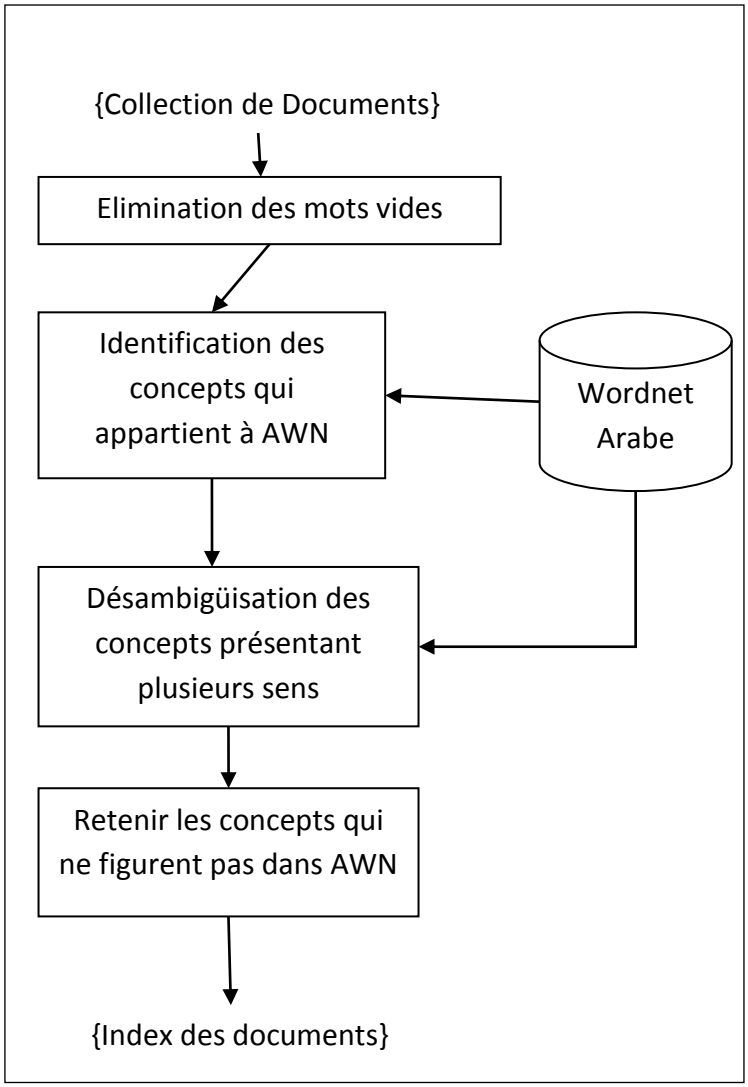


Figure 4.6 : IS des documents

3. Validation de l'approche proposée

L'évaluation se compose de deux partie, la première évaluation a pour but de mesurer l'apport de l'utilisation de l'ontologie (AWN) dans un SRI pour les textes arabes, la seconde partie sert à tester l'efficacité et évaluer l'apport de l'intégration de l'algorithme de lesk pour la désambigüisation du sens dans un SRI pour les textes arabes.

3.1. L'évaluation de l'apport de l'indexation sémantique

3.1.1. Expérimentation

Cette expérimentation est publiée dans le journal IJCSI [Abderrahim et al., 2013]. Elle a pour but de répondre à la question suivante : Quel est l'apport de l'intégration des ontologies dans un SRI pour les textes arabes ?

Afin de répondre à cette question nous avons effectué une expérimentation avec quatre types de recherche :

- Recherche Simple ou la recherche sans IS (R0): nous avons utilisé une liste de 70 requêtes simples de type mots-clés avec une indexation simple des documents.
- Recherche sémantique (R1) : nous avons indexé sémantiquement la collection des documents et la liste des 70 requêtes.
- Recherche avec expansion de la requête (R2) : nous avons indexé sémantiquement la liste des 70 requêtes. Pour la collection des documents, nous avons utilisé l'indexation classique (avec mots clés).
- Recherche avec IS des documents (R3) : nous avons indexé sémantiquement la collection des documents et nous avons utilisé la liste de 70 requêtes simples (mots-clés).

Dans le cadre de cette expérimentation, il est à noter que la désambiguïsation lors de l'IS est effectuée sur la base du concept commun (voir l'algorithme de désambiguïsation par le concept commun dans la section 2.2.1.1)

Nous avons calculé la précision dans les P@5, P@10, P@20, P@100, P@1000, premiers documents, et la précision moyenne.

Les résultats obtenus en termes de documents trouvés et de documents pertinents trouvés sont comme suit (voir tableau 4.5) :

Requête		Sans Indexation sémantique		Après indexation sémantique					
		R0		R1		R2		R3	
		Nb Doc Trouvés	Nb Doc Pertinents	Nb Doc Trouvés	Nb Doc Pertinents	Nb Doc Trouvés	Nb Doc Pertinents	Nb Doc Trouvés	Nb Doc Pertinents
1	ابداع	405	164	11588	6287	518	329	8937	6092
2	إثم	674	272	9332	5071	2579	1630	1914	1265
3	إدخار	366	96	4237	2225	3560	2163	357	95
4	استخدام	3539	361	17687	10985	9825	5564	3781	2438
5	اصابة	1940	889	3499	2034	2246	1511	1885	1384
6	اعداد	1632	258	17227	8475	9251	5046	3849	1596
7	إنتاج	1425	721	12892	6426	3496	2423	1389	927
8	انتهاك	237	71	4090	2445	2881	1598	226	147
9	بحث	8145	1016	19025	11373	11894	4837	13394	7206
10	تجهيز	264	41	4487	2612	4457	2612	254	82
11	تخفيض	553	386	11616	5828	3140	1410	545	386
12	تربية	3709	2297	14625	10158	6149	3963	4358	2713
13	تسامح	247	21	2977	1622	633	269	235	69
14	تصوير	408	113	15645	10355	7082	4531	720	453
15	تضخم	609	449	745	651	704	611	600	449
16	تطوير	1310	469	2399	1263	1858	600	1288	663
17	تعليم	2288	706	16678	10759	9505	7008	2828	1909
18	ثقب	92	39	6724	3375	1203	628	277	93
19	ثورة	1026	494	3130	2036	2441	1568	1088	526
20	جبهة	251	169	6783	3650	2102	1298	236	162
21	جرح	439	54	2890	2057	1394	983	1283	1030
22	جناية	237	95	572	357	604	380	213	82
23	جهد	1368	104	20903	12017	13622	7176	1334	787
24	حقنة	259	186	1725	1362	259	186	250	186
25	حماية	1299	383	12262	5807	8472	3818	1251	560
26	حياة	4214	1610	14912	9268	9926	5614	4909	3398
27	خلق	2369	876	14275	9441	7081	3701	3962	2051
28	دراسة	4451	1282	17486	10778	9524	5874	4407	2306
29	دنيا	2090	839	9000	6100	2090	1092	9000	6100
30	دين	7354	3067	12032	7262	11074	6766	7327	4307
31	سحابة	166	42	7479	1965	2465	1341	157	41
32	سعادة	1531	939	9982	6969	3562	2761	1507	934
33	سلوك	1841	640	7439	5990	1841	1231	7439	5990
34	صدق	1437	225	2929	2287	1437	1104	2929	2142
35	ضبط	1584	989	21742	15069	12752	9585	6111	4407
36	ضحك	261	88	574	395	261	197	574	395
37	طبخ	182	87	1818	731	182	87	1818	611
38	طعام	1856	713	9530	4913	5816	3546	5018	1341
39	طفل	1744	642	11987	8135	7799	4794	11413	3090
40	عائلة	1215	280	17708	14859	8427	4738	1186	766
41	عبادة	3032	574	8831	6142	3904	3411	2988	2697
42	عدوان	1244	587	6054	3182	4360	2603	1759	1005
43	علاج	3080	2046	6352	3676	4134	2024	3046	2044
44	عمل	6691	1652	22329	17791	19854	6953	14185	7832
45	عون	484	19	13632	8677	12724	8211	457	292
46	غذاء	1045	402	9530	4913	5816	2819	5565	2784
47	فريق	2304	1339	20345	14017	13825	7995	2376	1341
48	فضاء	681	423	6652	3161	4860	1414	663	423
49	فوز	1578	1129	6163	5267	1938	1154	3077	1451
50	قانون	1862	938	15532	11884	8687	1801	1832	937

Tableau 4.5 : Les documents trouvés et les documents pertinents trouvés pour chaque type d'indexation

Requête		Sans Indexation sémantique		Après indexation sémantique					
		R0		R1		R2		R3	
		Nb Doc Trouvés	Nb Doc Pertinents	Nb Doc Trouvés	Nb Doc Pertinents	Nb Doc Trouvés	Nb Doc Pertinents	Nb Doc Trouvés	Nb Doc Pertinents
51	قرض	320	132	2176	809	624	315	317	132
52	قصة	1524	413	8796	4738	4345	2423	1484	1232
53	قمر	697	410	773	413	697	410	773	413
54	كتابة	4482	1261	17199	9892	11289	3967	5670	3581
55	كذب	998	331	9886	5830	1887	1297	6163	3454
56	لاعب	1537	1379	8650	4786	2438	1393	5547	1776
57	لعب	1538	677	19703	11421	13085	5979	4776	2555
58	مالية	9095	2195	9332	5345	9115	5343	9179	2198
59	مجرة	835	352	942	354	835	352	812	351
60	مداواة	31	3	3541	1630	3335	443	28	3
61	مربي	2183	2026	4514	3195	2554	1730	2169	2016
62	مرض	2543	1043	4900	3427	2745	1049	4754	2231
63	مطالعة	114	40	5195	4696	5228	1506	111	89
64	معرفة	1696	545	19534	11706	8822	5835	1652	987
65	ملعب	529	494	1335	751	599	498	1286	649
66	مؤسسة	3071	1130	8463	5665	5695	2140	3043	1130
67	نادي	1458	1009	1828	1017	1864	1017	1436	1009
68	نجم	908	140	1416	868	908	606	1416	555
69	وظيفة	527	130	21932	12121	14077	5268	504	130
70	ولادة	170	50	7176	3071	573	297	155	49

Tableau 4.5 : Les documents trouvés et les documents pertinents trouvés pour chaque type d'indexation(Suite)

Une simple comparaison des résultats obtenus avant et après l'utilisation de la méthode d'IS pour représenter les documents et les requêtes, nous permet de déduire que cette méthode (pour tous les types) améliore dans la plupart des cas, le nombre de documents retournés et le nombre de documents pertinents retournés. En d'autres termes, l'IS améliore le rappel.

Concrètement soient :

- NDTB = Le nombre de documents trouvés avant la méthode d'IS.
- NDTA = Le nombre de documents trouvés après la méthode d'IS.
- D = NDTA - NDTB (1)
- NDTPB = Le nombre de documents pertinents trouvés avant la méthode d'IS.
- NDTPA = Le nombre de documents pertinents trouvés après la méthode d'IS.
- DP = NDTPA - NDTPB (2)
- Si (D> 0 ou DP> 0), alors on peut dire que l'IS améliore les performances de SRI en termes de rappel.
- En revanche, si (D = 0 ou DP = 0), en d'autres termes, nous avons le même nombre de documents retournés après l'IS. Donc, nous pouvons dire qu'il n'y a pas d'amélioration de la qualité de SRI d'un point de vue rappel.

Le tableau suivant est établi en se basant sur le calcul du nombre de requêtes en termes de D et DP :

	Documents Trouvés					
	Total des requêtes (R1)		Total des requêtes (R2)		Total des requêtes (R3)	
D<0	0	0%	0	0%	35	50%
D=0	0	0%	9	12.85%	0	0%
D>0	70	100%	61	87.15%	35	50%
	Documents Pertinents Trouvés					
	Total des requêtes (R1)		Total des requêtes (R2)		Total des requêtes (R3)	
DP<0	0	0%	13	18.6%	32	45.7%
DP=0	0	0%	3	4.3%	3	4.3%
DP>0	70	100%	54	77.1%	26	50%

Tableau 4.6 : La contribution d'IS basée sur les documents trouvés et les documents pertinents trouvés

Comme il est illustré dans tableau 4.6, l'augmentation du nombre de documents trouvés et les documents pertinents trouvés couvre pratiquement toutes les requêtes de R1. En outre, R2 et R3 sont les méthodes les moins appropriées pour l'IS (D <0) et (DP <0), car l'utilisation de la méthode d'IS modifie le vocabulaire soit dans les documents (R3) ou les requêtes uniquement (R2). Par exemple : le terme «إثم» qu'elle a remplacé dans l'indice sémantique de corpus par «خطيئة» et si nous cherchons en utilisant ce terme requête «إثم», le résultat sera négatif.

Selon le tableau 4.5, nous avons établi une comparaison entre les trois types de recherche (R1, R2 et R3) afin d'identifier la meilleure méthode d'IS du point de vue documents trouvés et documents pertinents trouvés. Le tableau suivant (voir tableau 4.7) présente les résultats de cette comparaison.

Documents Trouvés			
Pourcentage des requêtes dont laquelle R1 a envoyé plus de documents que les autres systèmes	Pourcentage des requêtes dont laquelle R2 a envoyé plus de documents que les autres systèmes	Pourcentage des requêtes dont laquelle R3 a envoyé plus de documents que les autres systèmes	Pourcentage des requêtes dont les trois systèmes (R1, R2, R3) ont envoyé le même nombre de documents
77.2%	1.4%	10%	0%
Documents Pertinents trouvés			
Pourcentage des requêtes dont laquelle R1 a envoyé plus de documents pertinents que les autres systèmes	Pourcentage des requêtes dont laquelle R2 a envoyé plus de documents pertinents que les autres systèmes	Pourcentage des requêtes dont laquelle R3 a envoyé plus de documents pertinents que les autres systèmes	Pourcentage des requêtes dont les trois systèmes (R1, R2, R3) ont envoyé le même nombre de documents pertinents
100%	0%	2.8%	0%

Tableau 4.7 : Comparaison entre les différents types de recherche (R1, R2 et R3)

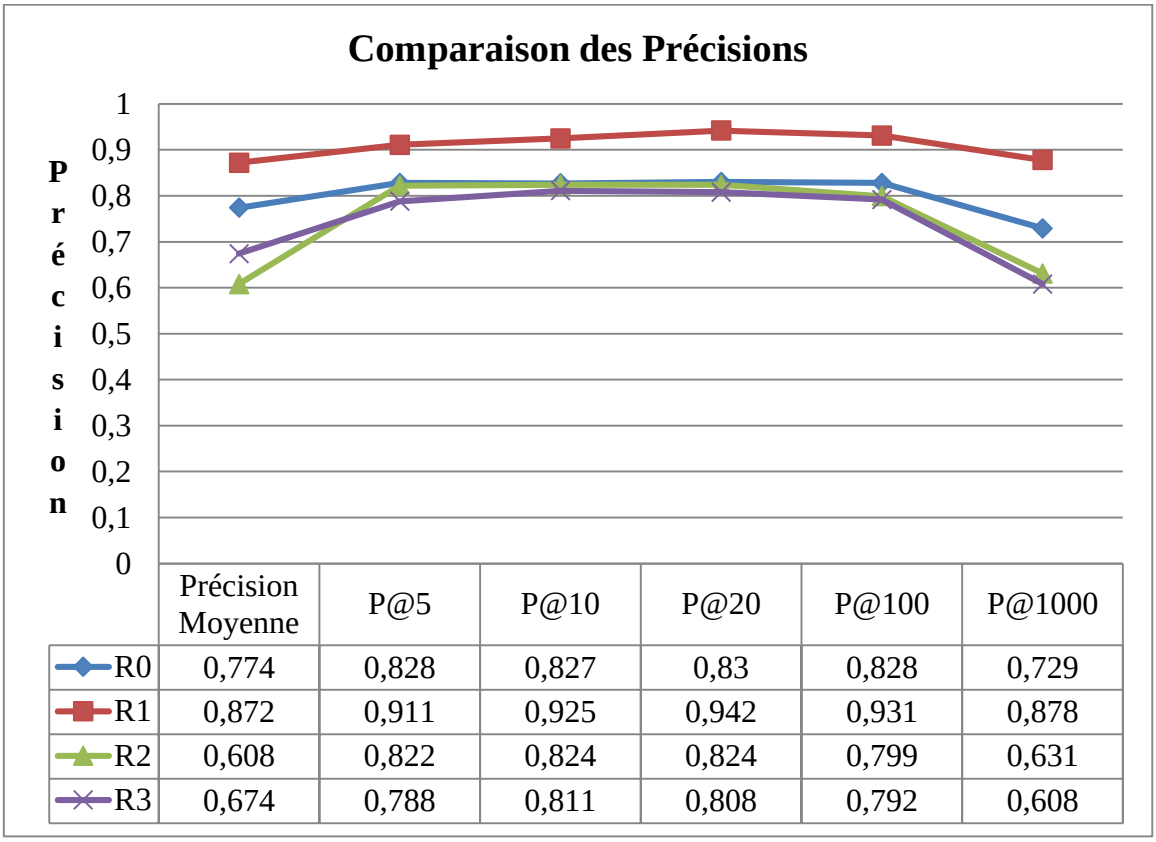
Les résultats présentés dans le tableau 4.7 donnent l'avantage au système R1 et par conséquent on peut dire que l'IS des documents et des requêtes présente ainsi le meilleur système de recherche d'un point de vue du nombre de documents trouvés et le nombre de documents pertinents trouvés. Ce résultat confirme le premier résultat qui a été déduit du tableau 4.6. D'un point de vue précision, le tableau 4.8 présente les différentes valeurs de précision obtenues dans les deux systèmes avant et après l'utilisation de la méthode d'IS.

		Précision Moyenne	**P@5**	**P@10**	**P@20**	**P@100**	**P@1000**
Avant Indexation Sémantique	R0	0,774	0,828	0,827	0,830	0,828	0,729
Après Indexation Sémantique	R1	**0,872**	**0,911**	**0,925**	**0,942**	**0,931**	**0,878**
	R2	0,608	0,822	0,824	0,824	0,799	0,631
	R3	0,674	0,788	0,811	0,808	0,792	0,608

Tableau 4.8 : Les différentes valeurs de précision obtenues par les deux systèmes

La comparaison des trois expérimentations par le graphique ci-dessous (voir Figure 4.7) nous montre que la méthode d'IS des documents et des requêtes (R1) donne le meilleur taux

de précisions dans toutes les valeurs de précision (P@5, P@10, P@20, P@100, P@1000), y compris la précision moyenne. Alors que, l'IS des documents et des requêtes séparément (R2, R3) donne des valeurs de précision moins bonnes pour toutes les mesures considérées.



	Précision Moyenne	P@5	P@10	P@20	P@100	P@1000
R0	0,774	0,828	0,827	0,83	0,828	0,729
R1	0,872	0,911	0,925	0,942	0,931	0,878
R2	0,608	0,822	0,824	0,824	0,799	0,631
R3	0,674	0,788	0,811	0,808	0,792	0,608

Figure 4.7 : Comparaison des valeurs de précisions des différents systèmes

3.1.2. Discussion

Du point de vue des documents trouvés et documents pertinents trouvés, nous pouvons dire que l'utilisation de la méthode d'IS pour représenter les deux documents et des requêtes ainsi améliore les performances d'un SRI. Du point de vue de la précision, (R1) à des bonnes valeurs pour toutes les mesures considérées ; par conséquent, il peut être choisi en tant que méthode pour représenter (indexation) des informations dans un SRI.

Si nous devons classer les autres méthodes (R2) et (R3), nous pouvons dire que R2 a l'avantage d'être plus précise pour les 5 et 10 et 20 et 100 et 1000 premiers documents, par contre, il présente de faibles valeurs pour la précision moyenne par rapport à R3.

3.2. L'évaluation de l'apport de l'indexation sémantique basée sur Lesk

3.2.1. Expérimentation

Cette expérimentation est publiée dans le journal International Journal of Speech Technology [Abderrahim et al., 2015]. Elle a pour but de répondre à la question suivante : Quel est l'apport de l'algorithme de désambigüisation (Lesk) s'il est utilisé dans l'IS des documents et des requêtes dans un SRI pour les textes arabes ? Autrement, peut-on améliorer les performances d'un SRI par l'utilisation des algorithmes de désambiguïsation plus performants que celui que nous avons utilisé dans l'expérimentation précédente ?

Afin d'évaluer l'approche d'IS nous avons procédé à trois types de recherche différents que nous allons les étudier séparément pour mesurer l'apport de chaque type dans l'amélioration des performances du SRI. Ces types de recherche sont :

– Recherche avec indexation simple (R0) : nous avons utilisé une liste de 50 requêtes simples de type mots clés. Une indexation par mots-clés des documents et des requêtes a été effectuée.

– Recherche avec IS et désambiguïsation par l'algorithme de Lesk (R1) : nous avons indexé sémantiquement la liste des 50 requêtes ainsi que la collection des documents utilisée. L'algorithme de Lesk est utilisé pour désambigüiser les termes ambigus.

– Recherche avec IS et désambiguïsation par la tête du Synset (R2) : nous avons indexé sémantiquement la liste des 50 requêtes ainsi que la collection des documents utilisée. Nous avons sélectionné le premier terme du Synset (tête du Synset) pour désambigüiser les termes ambigus.

Le tableau 4.9 présente les précisions à 11 points de rappels des différents systèmes (R0, R1, R2) associés à chaque type de recherche.

Rappel	Précision (R0)	Précision (R1)	Précision (R2)
0	0,731	1	0,815
0,1	0,641	0,877	0,787
0,2	0,6	0,77	0,725
0,3	0,495	0,757	0,695
0,4	0,417	0,628	0,645
0,5	0,4	0,615	0,541
0,6	0,377	0,535	0,477
0,7	0,347	0,453	0,446
0,8	0,3	0,437	0,368
0,9	0,188	0,317	0,252
1	0,117	0,209	0,179

Tableau 4.9 : Les précisions à 11 points de rappels selon le type de recherche

Avant de discuter les résultats obtenus, nous soulignons le fait que l'IS consomme beaucoup du temps machine comparé à l'indexation par mots-clés. En effet le tableau 4.10 donne les temps[27] machines consommées par l'opération d'indexation relative à chaque type de recherche.

	R0	R1	R2
Temps d'indexation de la collection des documents	4 mn et 36 s	23 h, 32mn et 12 s	18 h, 35 mn et 15 s
Taille de l'index généré en (M.O)	82.1	85	88.7

Tableau 4.10 : Les temps consommés par l'opération d'indexation relative à chaque type de recherche

D'après le tableau 4.10, nous constatons que l'IS consomme un temps très important (voir des heures) comparé à l'indexation par mots-clés (quelques minutes) ce qui a notre avis pose beaucoup de questionnement si nous envisageons son déploiement dans une application de grandeur réelle (comme l'IS du web).

La comparaison des valeurs de la précision dans les 11 points de rappels (Voir tableau 4.9) et des courbes rappel/précision (voir figure 4.8) pour les trois systèmes de recherche R0,

[27] Ce temps est obtenu par une machine de marque HP Core i5, 4 GO de Ram, 500 GO disque dur

R1 et R2 nous permet de déduire que l'IS de la requête et des documents améliore les performances d'un SRI pour les textes Arabes puisque les résultats du système R1 et R2 sont nettement meilleurs que ceux de R0. Nous remarquons aussi que les résultats obtenus par le système R1 sont plus performant que ceux du système R2 et sachant que ces deux systèmes (R1 et R2) sont similaires et ne se différencie que par l'algorithme utilisé pour la désambiguïsation des termes, nous pouvons alors déduire que l'algorithme de Lesk est plus performant que l'algorithme qui consiste à sélectionner le premier terme du Synset pour la désambiguïsation. Si nous avons donc à choisir entre l'algorithme utilisé dans R1 et celui utilisé dans R2, notre choix porte certainement sur celui de R1, néanmoins si nous prenons en compte le facteur temps d'exécution nous remarquons que l'algorithme de R2 est plus rapide que R1 (voir tableau 4.10) ce qui va naturellement justifier le choix de l'algorithme de R2 comme meilleur algorithme.

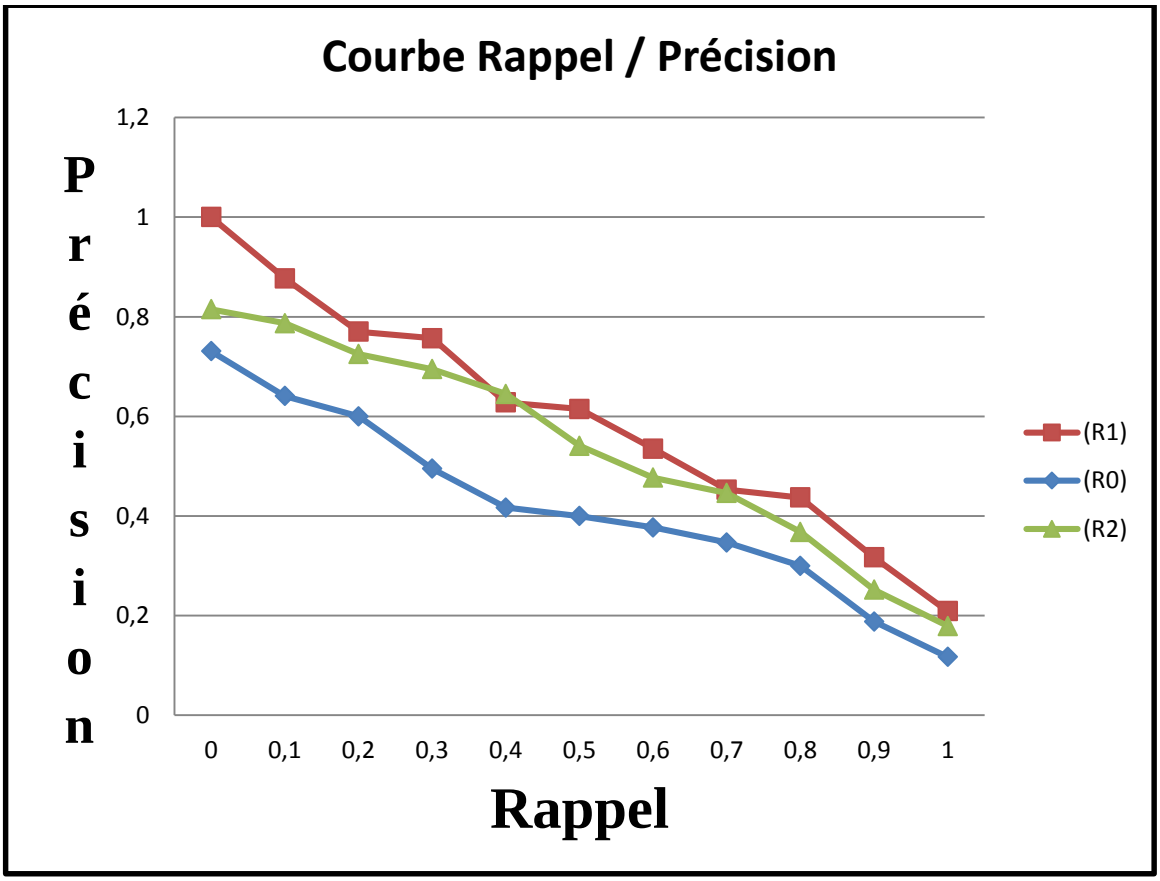


Figure 4.8 : Courbes rappels/précision selon le type de recherche

3.2.2. Discussion

L'expérimentation réalisée avait pour but d'évaluer l'approche d'IS avec et sans désambigüisation (par l'algorithme de Lesk) des termes utilisés dans le processus d'indexation des documents et requêtes pour les textes arabes. Nous avons commencé par l'IS de la collection des documents, qui été considéré comme une étape de préparation à la recherche, en utilisant une ressource sémantique (WordNet arabe dans notre cas). Puis, nous avons testé deux stratégies différentes de recherche (R1 et R2) qui reposent sur l'IS des documents et des requêtes. La comparaison, en termes de Rappel/Précision, des deux systèmes (R1 et R2) avec le SRI (R0) reposant sur une indexation par mots-clés nous a permet de conclure que l'IS est bien meilleur que l'indexation par mots-clés. Les résultats obtenues nous ont permet aussi de déduire que l'algorithme de Lesk reste le meilleur algorithme pour faire la désambigüisation de point de vue amélioration des performances (Rappel/Précision) d'un SRI pour les textes Arabes. Une conséquence directe est que nous pouvons dire que la désambiguïsation peut améliorer les performances d'un SRI pour les textes arabes.

4. Conclusion

Dans ce chapitre nous avons évalué l'approche d'IS pour les textes arabes, pour ce faire nous avons exploité la base lexicale WordNet arabe dans un SRI pour indexer la collection de documents et la requête utilisateur. Nos expérimentations effectuées sur un corpus arabe de taille moyenne nous ont montré que les ressources sémantique (pour notre cas : WordNet arabe) améliorent considérablement la qualité d'un SRI arabe. En faisant abstraction des temps machine importants dédiés aux opérations d'IS des collections de documents, l'apport des ontologies aux systèmes de recherche d'information en arabe est sans doute très intéressant et certes, mais comme il exige des ressources lexicales complètes, non disponibles à l'heure actuelle, il est encore tôt de notre point de vue, d'estimer avec certitude le taux de cet apport dans l'amélioration effectives des résultats de la recherche d'information.

Conclusion générale

La recherche d'information a pour objectif de fournir à un utilisateur un accès facile à l'information qui l'intéresse, cette information étant située dans une masse de documents textuels. Afin d'atteindre cet objectif, un système de recherche d'information doit représenter, stocker et organiser l'information puis fournir à l'utilisateur les éléments correspondant au besoin d'information exprimé par sa requête. Notre thèse s'inscrit dans le cadre de la recherche d'information pour les textes en langue arabe.

Ainsi un SRI pour les textes en langue arabe doit prendre en considération ses caractéristiques singulières et proposer des outils et des techniques automatiques afin de permettre son traitement informatique. L'objectif de notre travail a été d'une part, d'intégrer les ontologies (Wordnet Arabe) dans un système de recherche d'information, afin de passer d'une représentation par sac des mots à une représentation par des concepts.

Les travaux présentés dans cette thèse se situent dans le contexte général de l'utilisation de la sémantique pour la représentation de l'information dans la partie indexation des documents et des requêtes.

Une ressource sémantique externe est une structure pré-ordonnée qui peut se présenter sous forme d'ontologie, de hiérarchie de concepts, de réseau sémantique ou de thésaurus étendu. Le point commun à ses structures est qu'elles contiennent toutes un ensemble de concepts et de relations dénotant des liens sémantiques entre ces concepts. Le but est alors d'exploiter la sémantique contenue dans ces ressources, tout d'abord, pour une meilleure représentation de l'information et du besoin en information, puis, pour améliorer la correspondance entre le besoin de l'utilisateur et l'information. Dans le cadre des travaux présentés dans cette thèse nous avons exploité Wordnet Arabe, une ontologie légère pour la langue arabe librement téléchargeable sur le web.

Nous nous sommes intéressés dans cette thèse à proposer des solutions permettant de répondre à la question est-ce que l'introduction de l'information sémantique en RI est un moyen d'améliorer les performances d'un système de RI ?

Dans ce cadre, nous avons présenté principalement une contribution traduisant le point de vue de l'utilisation des ontologies en RI, à savoir, utiliser une ontologie (Wordnet Arabe) dans le processus d'indexation des documents et des requêtes. Elle sert dans ce cas, d'espace de

représentation conceptuelle dans lequel les documents et les requêtes sont exprimés par rapport à un référentiel commun. Cette approche vise à représenter l'information et le besoin en information non pas par rapport aux mots qu'ils contiennent mais par rapport aux concepts de l'ontologie auxquels ils renvoient.

Nous avons évalué l'approche d'indexation sémantique pour les textes arabe, cette évaluation se compose de deux partie, la première évaluation a pour but de mesurer l'apport de l'utilisation de l'ontologie (Wordnet Arabe) dans un SRI pour les textes arabes, la seconde partie sert à tester l'efficacité et évaluer l'apport de l'intégration de l'algorithme de lesk pour la désambigüisation du sens dans un SRI pour les textes arabes.

Nos expérimentations effectuées sur un corpus arabe de taille moyenne nous ont montré que les ressources sémantique (pour notre cas : WordNet arabe) améliorent considérablement la qualité d'un SRI arabe. En faisant abstraction des temps machine importants dédiés aux opérations d'indexation sémantique des collections de documents, l'apport des ontologies aux systèmes de recherche d'information en arabe est sans doute très intéressant et certes, mais comme il exige des ressources lexicales complètes, non disponibles à l'heure actuelle, il est encore tôt de notre point de vue, d'estimer avec certitude le taux de cet apport dans l'amélioration effectives des résultats de la recherche d'information.

Perspectives

Notre travail ouvre plusieurs perspectives :

- Etudier l'effet des autres méthodes de désambigüisations telle que [Agirre et al., 09] sur les performances des systèmes de recherche d'informations en arabe.
- Développer un système performant pour la désambigüisation des textes arabe.
- Améliorer la méthode de détection des termes d'indexation en se basant sur les autres relations existantes dans Wordnet Arabe.
- Proposer un modèle pour la pondération des concepts détectés dans Wordnet Arabe.
- Améliorer la couverture de WordNet Arabe.

Références bibliographiques

[Abderrahim, 08] Abderrahim Med El-Amine, Reconnaissance des unités linguistiques signifiantes, thèse de doctorat en informatique, université de tlemcen, 2008

[Abderrahim, 09a] Abderrahim M. A. : Vers une interface pour l'enrichissement des requêtes en arabe dans un système de recherche d'information. 2ème Conférence Internationale sur l'Informatique et ses Applications CIIA'2009, Saida - Algérie, 2009a.

[Abderrahim, 09b] Abderrahim, M.E.A.: "vers la recherche d'information de contenus en arabe fondée sur l'enrichissement des requêtes. In: Proceedings of the 2nd Conférence Internationale, Systèmes d'Information et Intelligence Economique, SIIE2009 Hammamet – Tunisie, 12-14 Février, Proceedings IHE éditions, ISBN 9978-9973-868-21-3, pp. 598–607,2009b.

[Abderrahim, 09c] Abderrahim, M.E.A.: Vers une interface pour l'enrichissement des requêtes en arabe dans un système de recherche d'information. In: Proceedings of the 2nd Conférence Internationale sur l'Informatique et ses Applications (CIIA'09), Saida, Algeria, May 3-4, pp. 60–69, 2009c.

[Abderrahim et al.,10] Abderrahim, Med-El-Amine and Med-Alaeddine A. Using Arabic Wordnet for Query Expansion in Information Retrieval System. In IEEE The Third International Conference on Web and Information Technologies, Marrakech, Morocco, June 2010.

[Abderrahim et al., 12] Abderrahim, Med-El-Amine and Med-Alaeddine A. Réinjection automatique de la pertinence pour la recherche d'informations dans les textes arabes. In IEEE 4th International Conference on Arabic Language Processing (CITALA), pages 77–81, Rabat, Morocco, May 2012.

[Abderrahim, 13a] Abderrahim, M.E.A. : Query Reformulation Guided by External Resource for Information Retrieval. World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:7, No:4,p 241-245, 2013a.

[Abderrahim, 13b] Abderrahim, M.A. : Utilisation des ressources externes pour la reformulation des requêtes dans un système de recherche d'information. In: the PBML 99 (The Prague Bulletin of Mathematical Linguistics), pp.87-99, 2013b.

[Abderrahim et al., 13] Abderrahim, Med-Alaeddine, A Med-El-Amine, and C Med-Amine.. Using Arabic WordNet for Semantic Indexation in Information Retrieval System. International Journal of Computer Science Issues, 10(2) : p 327–332. 2013.

[Abderrahim et al., 15] Mohammed Alaeddine Abderrahim, Mohammed Dib, Mohammed El-Amine Abderrahim & Mohammed Amine Chikh.. Semantic indexing of Arabic texts for information retrieval system. International Journal of Speech Technology. ISSN 1381-2416, p 1-8, 2015.

[Abouenour et al., 13] Abouenour Lahsen, Bouzoubaa Karim, Rosso Paolo, On the evaluation and improvement of Arabic WordNet coverage and usability. Volume 47, Issue 3, (pp 891-917). Springer September, 2013.

[Achour et al., 13] Achour, H., Zouari, M. Multilingual learning objects indexing and retrieving based on ontologies. World Congress on IEEE 2013 of the Computer and Information Technology (WCCIT). 2013.

[Agirre et al., 09] Agirre E, Soroa A. Personalizing PageRank for Word Sense Disambiguation. Proceedings of the 12th Conference of the European Chapter of the ACL, pages 33–41, Athens, Greece, 30 March – 3 April 2009. ©2009 Association for Computational Linguistics, 2009.

[Ahmed et Nürnberger, 08] Ahmed, F., Nürnberger, A.: Arasearch: Improving Arabic Text Retrieval via Detection of Word Form Variations. In: SIIE 2008 Hammamet – TunisieFévrier 14-16, pp. 309–323 2008.

[Aufaure et al., 07] Aufaure M. A., Soussi R., Baazaoui H., « SIRO: On-line semantic information retrieval using ontologies ». 2nd International Conference on Digital Information Management, ICDIM'07, p. 321- 326, 2007.

[Azzoug et al., 11] Azzoug W. Boubekeur F. Boughanem M. Indexation Sémantique de documents textuels. CIDE'11 : 14 ème Conférence Internationale sue le Document Electronique. Rabat, Maroc,2011.

[Azzoug et al., 12] Azzoug W. Boubekeur F. Boughanem M. Les concepts sont-ils de bons candidats à l'indexation ?. COSI'12: 9ème édition du colloque sur l'optimisation et les systèmes d'information. Tlemcen, Algérie, 2012.

[Azzoug, 13] Azzoug wassila. Mémoire de magister en informatique, Université M'hamed Bougara-Boumerdes, 2013.

[Azzoug et al., 13a] Azzoug W. Boubekeur F. Pondération des Concepts en Indexation Sémantique. CORIA'13 : Dixième édition de la Conférence en Recherche d'Information et Applications. Neuchatel, Suisse, 2013a.

[Azzoug et al., 13b] Azzoug W. Boubekeur F. Désambiguisation des sens des mots-application en recherche d'information. Dans 7ème Journées scientifiques pour la présentation des travaux de recherches des domaines de l'information, INFODays' 2013. Université Hassiba BenBouali. chlef, Algérie, 2013b.

[Bachimont, 00] B. Bachimont. Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In J. Charlet et al. (eds), Ingénierie des Connaissances; Evolutions récentes et nouveaux défis, Eyrolles, pp. 305-323, 2000.

[Baeza-Yates et al., 99] Ricardo A. Baeza-Yates, Berthier A. Ribeiro-Neto. Modern Information Retrieval ACM Press / Addison-Wesley 1999.

[Bakhouche et al., 12] Bakhouche, A., Tlili-Guiassa, Y. Meaning representation for automatic indexing of Arabic texts. International Journal of Computer Science Issues (IJCSI), 9(6), 173–178. (2012).

[Balack et al., 06] Black, William, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. Introducing the Arabic WordNet Project. In Proceedings of the Third International WordNet Conference, (pp 295–300). 2006.

[Baloul, 03] Baloul, S. Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé, Thèse de doctorat, Université du Maine, Académie de Nantes, France. 2003.

[Banerjee et Pedersen, 03] S. Banerjee, T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI, Acapulco, Mexico), p. 805–810, 2003.

[Baziz, 02] M. Baziz, Application des Ontologies pour l'Expansion de Requêtes dans un Système de Recherche d'Informations. Rapport de DEA Informatique de l'Image et du Langage (2IL). Université Paul Sabatier et Institut National Polytechnique de Toulouse, 2002.

[Baziz et al., 04] Baziz, M, Boughanem M, and Aussenac-Gilles N. The use of Ontology for Semantic Representation of Documents. The 2nd Semantic Web and Information Retrieval Workshop (SWIR), SIGIR 2004. Sheffield UK,Ying Ding, Keith Van Riejsbergen, Iad Ounis, Joemon Jose(Eds.), p 38-45, 2004.

[Baziz, 05] M. Baziz, Indexation conceptuelle guidée par ontologie pour la recherche d'information. Thèse de doctorat, université Paul Sabatier, 2005.

[Baziz et al., 05] M. Baziz, Boughanem M., Aussenac-Gilles N., Chrisment C., Semantic Cores for Representing Documents in IR, Proceedings of the 20th ACM Symposium on Applied Computing, pp 1020-1026, ACM Press ISBN: 1-58113-964-0, 2005.

[Belkin et al., 92] N. J. Belkin, W. B. Croft, Information retrieval and information filtering: Two sides of the same coin?. CACM, pages : 29-38, 1992.

[Benjamins et al., 99] Benjamins R., Fensel D., Decker D., Gomez Perez A., (KA)2, building ontologies for the internet : amid-term report, Proceedings of the International Workshop on Ontological Engineering on the Global Information Infrastructure, pp 1-24, (1999).

[Borst, 97] Borst, W. N. Construction of Engineering Ontologies. University of Tweenty. Enschede, NL- Centre for Telematica and Information Technology. 1997.

[Boubekeur, 08] F. Boubekeur. Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets. Thèse de doctorat en informatique, Université Toulouse III - Paul Sabatier, 2008.

[Boubekeur el al., 08] Boubekeur F. Boughanem M. Tamine L. Une approche d'indexation conceptuelle de documents basée sur les graphes CP_Nets. COSI'08: Cinquième édition du colloque sur l'optimisation et les systèmes d'information. Tizi-Ouzou, Algérie, 2008.

[Boubekeur et al., 10a] F. Boubekeur, M. Boughanem, L. Tamine , M. Daoud. Using WordNet for Concept- based document indexing in information retrieval. Fourth International Conference on Semantic Processing (SEMAPRO), Florence, Italy, October 2010.

[Boubekeur el al., 10b] F. Boubekeur, M. Boughanem, L.Tamine, M. Daoud. De l'utilisation de WordNet pour l'indexation conceptuelle des documents. le 13 ème Colloque International sur le Document Electronique (CIDE 13), 16-17 Décembre 2010, INHA, Paris, 2010.

[Boughanem et al., 92] Boughanem M. Soulé-Dupuy C. A Connexionist Model for Information Retrieval. DEXA 1992. p 260-265, 1992.

[Boughanem, 92] M. Boughanem : "les Systèmes de Recherche d'Information: d'un modèle classique à un modèle connexionniste", Thèse de Doctorat de l'Université Paul Sabatier, Toulouse (France), Décembre 1992.

[Boughanem, 00] M. Boughanem : Contribution à la Formalisation et à la Spécification des Systèmes de Recherche et de Filtrage d'Information. Habilitation à Diriger les Recherches, Université Paul Sabatier de Toulouse, 2000.

[Boughanem et al., 10] Boughanem M . Mallak I. Prade H. A new factor for computing the relevance of a document to a query. WCCI'10 : IEEE World Congress on Comutational Intelligence. Barcelone, 2010.

[Buitelaar et al., 07] P. Buitelaar, B. Magnini, C. Strapparava, P. Vossen. Domain specific word sense disambiguation, chapter 10. In Word sense disambiguation: algorithms and applications, p.275–298, 2007.

[Chang et al., 07] Chang B., Ham D.H., Moon D. S., Choi Y. S., Cha J., «Using Ontologies to Search Learning Resources». Book Series Lecture Notes in Computer Science, Éditeur Springer Berlin / Heidelberg ISSN 0302-9743, Book Computational Science and Its Applications – ICCSA, p. 1146-1159. Subject Collection Computer Science, 2007.

[Cowie et al., 92] J. Cowie, J. Guthrie, L. Guthrie. Lexical Disambiguation using Simulated Annealing. In Proceedings of the 14th International Conference on Computational Linguistics (COLING'92), Nantes, France, p. 359-365, 1992.

[Cucchiarelli et al., 04] Cucchiarelli R., Navigli R., Neri F., Velardi F., Extending and Enriching WordNet with OntoLearn, Proceedings of the 2nd Global WordNet Conference, 2004.

[Debili et Achour, 98] Debili F., Achour, H. Voyellation automatique de l'Arabe, Proceeding Semitic '98 Proceedings of the Workshop on Computational Approaches to Semitic Languages, 1998.

[Dilekh, 11] Tahar DILEKH. Implémentation d'un outil d'indexation et de recherche des textes en arabe. Mémoire de magister, Université Hadj Lakhdar – Batna, 2011.

[Dinh et al., 10] Dinh D. Tamine L. Vers un modèle d'indexation sémantique adapté aux dossiers medicaux de patients. CORIA'10: Conférence francophone en Recherche d'Information et Applications. P 325-336, 2010.

[Dinh, 12] Dinh D. Accés à l'information biomédicale : vers une approche d'indexation et de recherche d'information conceptuelle basée sur la fusion de ressources termino-ontologiques. Thèse Phd. Université de Toulouse 3 Paul Sabatier, 2012.

[Domingue, 99] Domingue J. Motta E., A knowledge based news Server Supporting Ontology-Driven Story Enrichment and Knowledge retrieval. In Proceedings of EKAW'99 11th European Workshop on Knowledge Acquisition, Modelling and Management. Berlin: Springer Verlag, LNAI, 1999.

[Douzidia, 04] Douzidia F. S.. Résumé automatique de texte arabe, Mémoire de M.Sc en informatique Université de Montréal, Québec, 2004.

[Dumais, 94] S. Dumais : Latent Semantic Indexing (LSI), TREC3 report. In Proceedings of the 3rd Conference on Text Retrieval Conference, 1994.

[Efthimiadis, 96] E. N. Efthimiadis. Query Expansion. In M. E. Williams, editor, Annual Review of Information Science and Technology, volume 31, pages 121–187. American Society for Information Science, 1996.

[Egozi et al., 11] O. Egozi, S. Markovitch, E. Gabrilovich. Concept-Based Information Retrieval using Explicit Semantic Analysis. ACM Transactions on Information Systems, Vol. 29 Issue 2, April 2011.

[Elketab et al., 06a] Elkateb, Sabry, William Black, H Rodríguez, M Alkhalifa, Piek Vossen, A Pease, and Christiane Fellbaum. Building a WordNet for Arabic. In Proceedings of The fifth international conference on Language Resources and Evaluation, Genoa-Italy, (pp 29–34).2006a.

[Elketab et al., 06b] Elkateb, Sabry, William Black, Piek Vossen, David Farwell, H Rodríguez, A Pease, and M Alkhalifa. Arabic WordNet and the Challenges of Arabic. In Proceedings of Arabic NLP/MT Conference, London, UK, p 15–24. Citeseer 2006b.

[Farquhar et al., 97] A. Farquhar, R. Fikes, and J. Rice : The ontoloingua server : A tool for collaborative ontology construction. Journal of Human-Computer Studies, 46 :707-728, 1997.

[Fensel, 01] Fensel D. Ontologies: a silver bullet for Knowledge Management and Electronic Commerce. Berlin : Springer Verlag, 2001.

[Fox, 83] E. Fox, Extending the boolean and vector space models of information retrieval with p-Norm queries and multiple concept types, PhD Thesis Cornell University, 1983.

[Fürst, 04] F. FÜRST, Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationnalisation. Thèse de doctorat, École Polytechnique de l'Université de Nantes (EPUN) 2004.

[Gasmi, 09] Gasmi, M. Utilisation des ontologies pour l'indexation automatique des sites Web en Arabe. Mémoire de magister, Université Kasdi Merbah Ouargla. 2009.

[Gliozzo et al., 04] A. Gliozzo, B. Magnini, C. Strapparava. Unsupervised domain relevance estimation for word sense disambiguation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP, Barcelona, Spain), p.380–387, 2004.

[Gomez, 99] A.Gomez, « Développements récents en matière de conception, de maintenance et d'utilisation d'ontologies ». in 3èmes rencontres Terminologie et intelligence artificielle TIA 1999.

[Gruber, 93] T. R. Gruber, "Toward Principles for the design of Ontologies used for Knowledge Sharing," in Proc of International Workshop on Formal Ontology, Padova, Italy, March 1993.

[Guarino et al., 99] Guarino, N., C. Masolo, and G. Vetere, OntoSeek: Using Large Linguistic Ontologies for Accessing On-Line Yellow Pages and Product Catalogs,. National Research Council, LADSEBCNR: Padavo, Italy. 1999.

[Guha et al., 03] Guha R. V., McCool R., Miller E. : Semantic search, Proceedings of the 12th International World Wide Web Conference, pp 700-709, 2003.

[Guthrie et al., 91] J.A. Guthrie, L. Guthrie, Y. Wilks, H. Aidinejad. Subject-dependant cooccurrence and word sense disambiguation. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkley, p. 146-152, 1991.

[Haav et Lubi, 01] Haav H. M., Lubi T.L., A Survey of Concept-based Information Retrieval Tools on the Web, Proceedings of the 5th East-European Conference ADBIS, Vol 2, pp 29-41, 2001.

[Haines et al., 93] D. Haines, W. Bruce Croft: Relevance Feedback and Inference Networks. SIGIR : 2-11. 1993.

[Hammache et al., 09] A. Hammache, M. Boughanem, R. Ahmed-Ouamer. Introduction de la sémantique d'un document sous le modèle de langage. Dans la sixième édition de la Conférence francophone en Recherche d'Information et Applications (CORIA 2009), 5-7 mai 2009.

[Hammo et al., 07] Hammo, B., Sleit, A., El-Haj, M.: Effectiveness of Query Expansion in Searching the Holy Quran. In: colloque internationale Traitement automatique de la langue Arabe:, CITALA, vol. 7, pp. 18–19, 2007.

[Harman, 92] Donna Harman: Relevance Feedback and Other Query Modification Techniques. Information Retrieval: Data Structures & Algorithms: 241-263, 1992.

[Harrathi et al., 10] Harrathi F. Roussey C. Maisonnnasse L. Calabretto S. Vers une approche statistique pour l'indexation sémantique des documents multilingues. Actes du XXVIII° congrés INFORSID. Marseille, 2010.

[Hearst et Karadi, 97] Hearst M.A, Karadi C., Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy, Proceedings of the 20th International conference on Research and Development in Information Retrieval, SIGIR, pp 246-257, 1997.

[Hearst, 97] M.A. Hearst, C. Karadi, Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy, In Proceedings of the 20th International conference on Research and Development in Information Retrieval, SIGIR, pp 246-257, 1997.

[Hernandez, 05] Hernandez, N. Ontologies de Domaine pour la Modélisation du Contexte en Recherche d'Information. Thèse Phd, Université Toulouse III-Paul Sabatier, 2005.

[Hernandez et al., 07] Hernandez N., Mothe J., Ralalason B., Ramamonjisoa B., Stolf P., « Multi-facet indexing for learning objects reuse ». Dans : Computer Science & Information Technology Education, Pointe-aux-Sables, 16/11/2007-18/11/2007, Institute for Scientific Information (ISI), p.309-322, 2007.

[Hernandez et al., 08] Hernandez N., Hubert G., Mothe J., Ralalason B. : RI et Ontologies - Etat de l'art 2008, RAPPORT INTERNE N° IRIT/RR-2008-14-FR, JUILLET 2008.

[Hlaoua, 07] L. Hlaoua : Reformulation de Requêtes par Réinjection de Pertinence dans les Documents Semi-Structurés. Thèse de doctorat, université Paul Sabatier (2007)

[Kanaan et al.,05] Kanaan, G., Al-Shalabi, R., Abu-Alrub, M., Rawashdeh, M.: Relevance Feedback: Experimenting with a Simple Arabic Information Retrieval System with Evaluation. International Journal of Applied Science and Computations Vol 12 No 2 USA, 2005.

[Karbasi, 07] S. Karbasi : Pondération des termes en Recherche d'Information: Modèle de pondération basé sur le rang des termes dans les documents. Thèse de doctorat, université Paul Sabatier, 2007.

[Kelly et al., 75] E. F. Kelly, P. J. Stone. Computer recognition of english word senses. North-Holland Publishing. North-Holland, Amsterdam, 1975.

[Khan, 00] Latifur R. Khan, Ontology-based Information Selection, Phd Thesis, Faculty of the Graduate School, University of Southern California. August 2000.

[Khan et al., 04] Khan L. R., Mc Leod D., Hovy E. Retrieval effectiveness of an ontology based model for information selection. The VLDB Journal, edition 13, p 71-85, 2004.

[Kim et al., 07] Kim H., Park C. S., Park J. Y., Jung B., Lee Y. J., « A Multimedia Content Management and Retrieval System Based on Metadata and Ontologies ». IEEE International Conference on Multimedia and Expo, pp. 556 – 559, 2007.

[Köhler et al., 06] Köhler J., Philippi S., Specht M., Rüegg A., « Ontology based text indexing and querying for the semantic web ». Knowledge-Based Systems, Vol 19, Issue 8, December 2006, p. 744 – 754, 2006.

[Kolar et al., 05] Kolar M.; Vukmirovic I., Basic B.D., Snajder J., « Computer aided document indexing system ». 27th International Conference on Information Technology Interfaces, 2005. p. 323 – 328, 2005.

[Kolte et al., 08] S.G. Kolte, S. G. Bhirud. Word Sense Disambiguation using WordNetDomains. In First International Conference on Emerging Trends in Engineering and Technology. IEEE DOI 10.1109/ICETET, 2008.

[Kolte et al., 09] S. G. Kolte, S. G. Bhirud. WordNet : A Knowledge Source for Word Sense Disambiguation. International Journal of Recent Trends in Engineering, Vol 2, No. 4, November 2009.

[Kompaoré, 08] Y. Kompaoré, Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes : vers un processus de RI adaptatif. Thèse de doctorat, université Paul Sabatier, 2005.

[Kouloughli, 94] Djamel Kouloughli. Grammaire de l'arabe d'aujourd'hui ; Pocket-Langues pour tous, 1994.

[Krovetz et al., 92] R. Krovetz, W.B. Croft . Lexical Ambiguity and Information Retrieval, in ACM Transactions on Information Systems, 10(1), 1992.

[Lancaster, 68] Lancaster, F.W., Evaluation of the MEDLARS Demand Search Service, National Library of Medicine, Bethesda, Maryland, 1968.

[Lenat et al., 90] D.B. Lena and R.V. Guha :Building large knowledge-based systems.Representation and inference in the Cyc project, Addison- Wesley, Reading,Massachusetts, USA, 1990.

[Lesk, 86] M.E. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a nice cream cone. In Proceedings of the SIGDOC Conference, Toronto, 1986.

[Luhn, 57] Luhn, H., A statistical approach to mechanized encoding and searching of literary information. IBM, 1(4):309–317, 1957.

[Magnini et al., 00] B. Magnini, G. Cavagli. Integrating subject field codes into WordNet. In Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, p. 1413- 1418, 2000.

[Maisonnasse et al., 09] Maisonnasse L. Gaussier E. Chevallet J-P. Model Fusion in Conceptual Language Modeling. ECIR 2009. P 240-251, 2009.

[Mallak, 11] Mallak I. De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en recherche d'information. Thèse Phd. Université de Toulouse, 2011.

[Masolo, 01] Masolo C. Ontology driven Information retrieval: Stato dell'arte. Report of the IKF (Information and Knowledge Fusion) Eureka Project E!2235. LADSEBCnr, Padova (I), 2001.

[Mihalcea et Moldovan, 00] Mihalcea, R. and Moldovan, D.: Semantic indexing using WordNet senses. In Proceedings of ACL Workshop on IR & NLP, Hong Kong, October 2000. http://www.seas.smu.edu/~rada/papers/acl00.nlp_ir.ps.gz

[Miller, 90] MILLER G.A., BECKWITH R., FELLBAUM C., GROSS D., MILLER K.: " Introduction to WordNet : An on-line lexical database ", Journal of Lexicography, n°3, pp.235-244, 1990.

[Mohammad et al., 06] S. Mohammad, G.Hirst. Determining word sense dominance using a thesaurus. In Proceedings of the 11th Conference on European chapter of the Association for Computational Linguistics (EACL, Trento, Italy), p. 121–128, 2006.

[Moldovan et al., 99] Moldovan D., Harabagiu S., Pasca M., Mihal-cea R., Goodrum R., Girju R., Rus V., LASSO: A tool for surfing the answer net. Proceedings of the 8th Text Retrieval Conference (TREU-8), 1999.

[Mooers, 48] C.N. Mooers. Application of random codes to the gathering of statistical information. MIT Master's Thesis. 1948.

[Mothe, 94] J. Mothe : Modèle Connexionniste pour la Recherche d'Information, Expansion dirigée de requêtes et apprentissage. Thèse de Doctorat, université Paul Sabatier. 1994.

[Nassr, 02] N. Nassr, Croisement de langues en recherche d'information: traduction et désambiguïsation de requêtes. Thèse de doctorat, université Paul Sabatier. 2002.

[Ponte et al., 98] Ponte, J. and Croft, W., A language modelling approach to information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 40–48, 1998.

[Pantel et al., 02] P. Pantel, D. Lin. Discovering word senses from text. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (Edmonton, Alta., Canada), p.613-619, 2002.

[Robertson, 77] S. E. Robertson: The probability ranking principle in IR. Journal of Documentation, 33 (4), 294-304, 1977.

[Robertson et Walker, 94] S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Proc. of the International ACM-SIGIR Conference(1994), pages 232–241,1994.

[Salton, 71] G. Salton : The Smart Retrieval System : Experiments in Automatic Document Processing, G. Salton Editor, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1971.

[Salton, 83] Salton, G., Introduction to modern information retrieval. New York, McGraw-Hill. 1983.

[Salton, 89] Salton, G., Automatic text processing : The transformation, analysis and retrieval of information by computer. Addison-Wesley publishing, MA. 1989.

[Schütze, 98] H. Schütze. Automatic word sense discrimination. Computational Linguistic: Special Issue on Word Sense Disambiguation, 24 (1), p.97–123, 1998.

[Singhal et al.,96] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. Information Processing and Management, 32(5) :619–633, 1996.

[Sussna, 93] M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. 2nd International Conference on Information and Knowledge Management (CIKM- 1993), p.67–74, 1993.

[Swartout et al., 97] Swartout (B.), Patil (R.), Knight (K.) and Russ (T.): Towards Distributed Use of Large-Scale Ontologies. Spring Symposium Series on Ontological Engineering, Stanford University, CA, p. 138-148. 1997.

[Tamine, 00] L. Tamine, Optimisation de requêtes dans un Système de Recherche d'Information. Thèse de doctorat, université Paul Sabatier. 2000.

[Tazzite et al., 08] Tazzite, N., Yousfi, A., Bouyakhf, H. Conception et réalisation d'un système de recherche d'informations intégrant des connaissances sémantiques dans la phase d'indexation. NTIC'08, Les Technologies de l'information: statuts ET opportunités pour l'amazighe. Rebat MAROC. Retrieved from 28 Nov 2008. 2008.

[Tebri, 04] H. Tebri : Formalisation et spécification d'un système de filtrage incrémental d'information. Thèse de doctorat, université Paul Sabatier. 2004.

[Tomassen et al., 06] Tomassen S. L., Gulla J. A., Strasunskas D., « Document Space Adapted Ontology: Application in Query Enrichment ». 11th International Conference on Applications of Natural Language to Information Systems. Springer, Klagenfurt, Austria, 2006.

[Turtle et al., 92] Howard R. Turtle, W. Bruce Croft: A Comparison of Text Retrieval Models. Comput. J. 35(3): 279-290. 1992.

[Turtle et Croft, 91] H. Turtle, W.B Croft : Evaluation of an Inference Network Based Retrieval Model. ACM Transactions on Information Systems July, 1991.

[Vallet, 05] D. Vallet, M. Fernández, P. Castells, An Ontology-Based Information Retrieval Model, In Proceedings of the 2nd European Semantic Web Conference, pp 455-470, 2005.

[Vallet et al., 07] Vallet, D, Castells P, Fernández M, Mylonas P, and Avrithis Y. Personalized content Retrieval in Context using Ontological Knowledge. IEEE Transactions on Circuits and Systems for Video Technology, n° 17. p 336–346, 2007.

[Vasilescu, 03] Vasilescu, F. Monolingual corpus disambiguation by the approaches of Lesk. Master's thesis. University of Montreal, Faculty of Arts and Sciences. 2003.

[Vázquez et al., 04] S. Vázquez, A. Montoyo, G. Rigau. Using Relevant Domains Ressource for Word Sense Disambiguation. Proceeding of international conference on Artificial intelligence, (IC- AI'04), Nivada, 2004.

[Véronis et al., 90] J. Véronis, N. Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. 13th International Conference on Computational Linguistics (COLING-1990), Vol.2, p.389–394. 1990.

[Voorhees, 93] E. Voorhees, "Using WordNet to Disambiguate Word Senses for Text Retrieval", Proceedings of the 16th Annual Conference on Research and Development in Information Retrieval, SIGIR'93, Pittsburgh, PA, 1993

[Voorhees, 94] Voorhees E. M., Query expansion using lexical-semantic relations, Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 61-69, 1994.

[Wedyan, 12] Wedyan, M., Alhadidi, B., Alrabea, A.: The effect of using a thesaurus in arabic information retrieval system. International Journal of Computer Science Issues, IJCSI 9(1), 431–435, 2012.

[Weiss, 73] S.F. Weiss. Learning to disambiguate, in Information Storage and Retrieval, 9: pp 33-41, 1973.

[Wilks et al., 90] Y. Wilks, D. Fass, C. Guo, J.E. Mcdonald, T. Plate, B.M. Slator. Providing Machine Tractable Dictionary Tools. In Machine Translation, Vol.5, p.99-154. 1990.

[Xiaomeng et Atle, 06] Xiaomeng S., Atle J. G., « An information retrieval approach to ontology mapping ». Data & Knowledge Engineering, Vol. 58 Issue 1, pp. 47-69, 2006.

[Xu et al.,02] Xu, J., Fraser, A., Weischedel, R.: Empirical Studies in Strategies for Arabic Retrieval. In: Annual ACM Conference on Research and Development in Information Retrieval: Proceedings of the 25 th annual international ACM SIGIR conference on Research and development in information retrieval, vol. 11, pp. 269–274, 2002.

[Yarowsky, 92] D. Yarowsky. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. Proceedings of the 14th International Conference on Computational Linguistics (COLING-92). Nantes, France, August, p.454 – 460, 1992.

[Yarowsky, 00] D. Yarowsky. Hierarchical decision lists for word sense disambiguation. Journal Computers and the Humanities, Vol.34 (1-2), p179-186, 2000.

[Zaidi et Laskri, 07] Zaidi, S., Laskri, M.: Expansion de la requête Arabe sur le réseau internet. In: Barmajiat (CSLA): Les applications logicielles en arabe: Pas vers le e-gouvernement 9-10 Décembre Alger, 2007.

[Zaidi, 13] Soraya Zaidi–AyadUne plateforme pour la construction d'ontologie en arabe : Extraction des termes et des relations à partir de textes (Application sur le Saint Coran), Thèse de doctorat, université, Badji Mokhtar, Annaba, 2013.

[Zhao et al., 07] Zhao Y., Zhang J., Guan B., Hu J., Wang W., « The Development of Intelligent Retrieval Algorithm Ontology-based And Its Application in Bearing production Information System ». 11th International Conference on Computer Supported Cooperative Work in Design, 2007. CSCWD'07, pp. 722 – 727, 2007.

[Zouaghi et al., 11] A. Zouaghi, L. Merhbene, and M. Zrigui, "Word sense disambiguation for arabic language using the variants of the lesk algorithm," in Proceedings of the International Conference on Artificial Intelligence (ICAI'11), vol. 2, pp. 561–567, 2011.

[Zouaghi et al., 12] Zouaghi A, Zrigui M, Antoniadis G, et Merhbene L: "Contribution to semantic analysis of Arabic language", Journal Advances in Artificial Intelligence, Volume 2012, Article No. 11, Hindawi Publishing Corp. New York, NY, United States, January 2012.

[Zweigenbaum, 93] Zweigenbaum P. et al., Linguistic and medical knowledge bases: An access system for medical records using natural language, Technical report, MENELAS: deliverable 9, AIM Project A2023, 1993.

Résumé

Les travaux présentés dans cette thèse se situent dans le contexte général de l'utilisation de la sémantique pour la représentation et la manipulation de l'information dans un système de recherche d'information pour les textes Arabes. L'objectif est alors d'exploiter la sémantique contenue dans des ressources sémantique, tout d'abord, pour une meilleure représentation de l'information et du besoin en information, puis, pour améliorer la correspondance entre le besoin de l'utilisateur et l'information.

Dans ce cadre, nous avons présenté principalement une contribution traduisant le point de vue de l'utilisation de l'ontologie WordNet Arabe (AWN) dans le processus d'indexation des documents et des requêtes. Par ailleurs, pour la validation nous avons évalué cette approche dans le but de mesurer l'apport de l'utilisation de AWN. En effet, nos expérimentations effectuées sur un corpus arabe de taille moyenne nous ont montré que les ressources sémantiques améliorent considérablement la qualité des systèmes de recherche d'information pour les textes arabes.

Mots clés : Système de recherche d'information, Ontologie, Indexation Sémantique, Désambiguïsation, WordNet Arabe (AWN).

Abstract

Investigations conducted in this thesis are in the general context of the use of semantics for the representation and manipulation of information in information retrieval system for Arabic texts. The aim is to use the semantics contained in the semantic resources, first, for a better representation of the information and the need of information, then, to improve the correspondence between the user needs and information.

In this context, we presented a contribution reflecting the use of Arabic WordNet (AWN) ontology in the process of indexing documents and queries. Furthermore, for the validation we evaluated this approach in order to measure the AWN contribution use. Our experiments performed on an Arabic corpus have shown us that the semantic resources significantly improve the quality of information retrieval systems for Arabic texts.

Keys words: Information retrieval system, Ontology, Semantic Indexing, Disambiguation, Arabic WordNet (AWN).

ملخص

تندرج الأبحاث في هذه الأطروحة تحت السياق العام لاستخدام المستوى الدلالي لتمثيل ومعالجة المعلومات في أنظمة البحث عن المعلومات الخاصة بالنصوص العربية . الهدف من ذلك هو استخدام الدلالة الموجودة في الموارد الدلالية، لأجل، التمثيل الأفضل للمعلومات والحاجة للمعلومات، من جهة، وتحسين مدى التوافق بين احتياجات المستخدم والمعلومات من جهة أخرى.

في هذا السياق، قدمنا مساهمة تعكس استخدام الأنطولوجيا العربية WordNet في عملية فهرسة الوثائق وطلبيات المستخدم. و من أجل التحقق من صحة فرضيتنا قمنا بتقييمها وهذا من أجل قياس مدى جدوى استخدام هذه الأنطولوجيا . لقد أظهرت لنا التجارب التي أجريت على مدونة للعربية أن الموارد الدلالية تحسن فعلا جودة أنظمة البحث عن المعلومات الخاصة بالنصوص العربية.

الكلمات المفتاحية: أنظمة البحث عن المعلومات ، الأنطولوجيا، الفهرسة الدلالية، حل اللبس، الأنطولوجيا العربية WordNet.