REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE : Mathématiques et Informatique
DEPARTEMENT : Informatique
N° :……..

DOMAINE : INFORMATIQUE
FILIERE : INFORMATIQUE
OPTION : SIGL

Mémoire présenté pour l'obtention
Du diplôme de Master Académique

Par: AZIZI Abdelali

Intitulé

CLASSIFICATION AUTOMATIQUE DE TEXTES ARABES SUPERVISEE PAR L'ONTOLOGIE LEXICALE WORDNET

Soutenu devant le jury composé de :

BOUDIA MALIKA	Université de M'sila	Président
KADRI SAID	Université de M'sila	Rapporteur
BOUDAA ABDELGHANI	Université de M'sila	Examinateur

Année universitaire : 2017 /2018

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE : Mathématiques et Informatique
DEPARTEMENT : Informatique
N° :……..

DOMAINE : INFORMATIQUE
FILIERE : INFORMATIQUE
OPTION : SIGL

Mémoire présenté pour l'obtention
Du diplôme de Master Académique

Par: AZIZI Abdelali

Intitulé

CLASSIFICATION AUTOMATIQUE DE TEXTES ARABES SUPERVISEE PAR L'ONTOLOGIE LEXICALE WORDNET

Soutenu devant le jury composé de :

BOUDIA MALIKA	Université de M'sila	Président
KADRI SAID	Université de M'sila	Rapporteur
BOUDAA ABDELGHANI	Université de M'sila	Examinateur

Année universitaire : 2017 /2018

TABLE DES MATIERES

CHAPITRE 2: LES ONTOLOGIES CONCEPTS ET NOTION DE BASES:

CHAPITRE 3: L'ONTOLOGIE LEXICALE WORD NET

CHAPITRE 4: IMPLIMENTATION ET EXPERIMANTATION

Listes des figures et tableaux

Figures :

Tableaux

Introduction Générale

La révolution de l'information bousculée par le développement à grande échelle des accès réseaux Internet/Intranet a fait exploser la quantité d'informations textuelles disponibles en ligne ou hors ligne et la vulgarisation de l'informatique dans le monde des entreprises, des administrations et des particuliers, a permis de créer des volumes importants de documents électroniques rédigés en langue naturelle. Il est très difficile d'estimer les quantités de données textuelles crées chaque mois dans les administrations, les sociétés, les institutions, ou la quantité de publications scientifiques dans les divers domaines de recherche.

L'information textuelle qui prend de plus en plus d'importance dans l'activité quotidienne des chercheurs et des entreprises ainsi que les besoins d'accès intelligents aux immenses bases de données textuelles et leurs manipulations qui ont augmenté très largement, d'une part. D'autre part les limites d'une approche manuelle qui est coûteuse en temps de travail, peu générique, et relativement peu efficace, ont motivé la recherche dans ce domaine.

Ainsi la recherche des solutions opérationnelles, et la mise en œuvre d'outils efficaces pour automatiser la classification de ces documents devient une nécessité absolue. De nombreux travaux de recherche se focalisent sur cet aspect donnant ainsi un nouvel élan à la recherche dans le domaine qui connait une évolution réelle depuis les deux dernières décennies. Comment partitionner cette masse d'information en groupes ou classes pour dégager des ressemblances par thèmes, par auteurs, par langue, ou par d'autres critères de classification ou carrément un filtrage de l'ensemble de documents utiles parmi les documents inutiles (Cas des filtres anti-spams).

C'est à ce niveau que se positionne notre problématique de classification de textes. L'objectif de la classification de textes est de rassembler les textes similaires selon un certain critère, au sein d'une même classe. Deux types d'approches de classification automatique peuvent être distingués : La classification supervisée et la classification non supervisée.

Ces deux méthodes diffèrent sur la façon dont les classes sont générées. En effet dans le cas de la classification non supervisée, les classes sont calculées automatiquement par la machine, par contre, dans l'approche supervisée, la classification de textes consiste à rattacher un texte à une ou plusieurs catégories prédéfinies par un expert, ces catégories pouvant être par exemple le sujet du texte, son thème, l'opinion qui y est exprimée, etc... Nous disposons pour cela d'un ensemble de textes pour lesquels la catégorie est connue (corpus

d’apprentissage) et qui nous servent à entraîner nos modèles, modèles qui seront testés et évalués sur d’autres documents pour lesquels la catégorie est connue également (corpus de test), le meilleur de ces modèles sera adopté par la suite pour étiqueter automatiquement des nouveaux documents de catégorie indéterminée. La problématique de classification nous conduit à nous placer dans l’intersection de plusieurs disciplines variées :

L’étude que nous avons mené sur les différentes approches de classification de textes conçues par un programme représentant l'expert qui est capable de résoudre le problème par lui-même ou concevoir des programmes comme des sortes de penseurs repliés sur eux-mêmes a trouvé sa limitation lorsque nous avons cherché à développer des modèles plus complexes de classification de bases de données textuelles gigantesques réalisées habituellement non pas par une seule personne mais par un groupe de personnes parfois délocalisées.

Ces limitations peuvent être ressenties facilement par une dégradation considérable des performances des meilleurs classifieurs, en temps de réponse qui s’augmente proportionnellement avec la taille des volumes traités et même pour la qualité des résultats.

Dans notre travail, nous avons adopté une approche basée sur l’utilisation d’une ressource sémantique fournissant l’information pertinente afin d’améliorer le processus de catégorisation, surtout dans sa phase de représentation. Il s’agit ici de l'ontologie lexicale WordNet qui sert à capturer toute relation sémantique entre les termes ce qui permet de réduire la dimension de l'espace de représentation des textes d'une part, et d'améliorer l'exactitude de la catégorisation proprement dite d'une autre part.

Notre mémoire est représenté comme suit :

- Une introduction décrivant le domaine de recherche et notre problématique.
- Chapitre 1 : Classification automatique des documents
- Chapitre 2 : Les ontologies concept et notion de base
- Chapitre 3 : L’ontologie l’lexicale WORD NET
- Chapitre 4 : La réalisation
- Une conclusion générale qui synthétise les résultats obtenus, les outils utilisés, les difficultés rencontrées, ainsi que les perspectives estimées.

CHAPITRE1
CLASSIFICATION AUTOMATIQUE DES DOCUMENTS

1- Introduction

La classification de textes est une tâche générique qui consiste à assigner une ou plusieurs catégories, parmi une liste prédéfinie, ou non à un document. Est un problème qui intéresse les chercheurs relativement longtemps. Actuellement, La recherche dans ce domaine est toujours très pertinente, car les résultats obtenus aujourd'hui sont encore sujets à amélioration malgré ont vu plusieurs modèles.

2- La Classification

2.1 Définition

La classification est la tâche la plus commune du Data Mining et qui semble être une obligation humaine. Afin de comprendre notre vie quotidienne, nous sommes constamment classifiés, catégorisés et évalués.

La classification consiste à étudier les caractéristiques d'un nouvel objet pour lui attribuer une classe prédéfinie. Les objets à classifiés sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jour chaque enregistrement en déterminant un champ de classe. La tâche de classification est caractérisée par une définition de classes bien précise et un ensemble d'exemples classés auparavant. [1]

2.2 But de la classification

Comme les autres méthodes de l'Analyse des données, dont elle fait partie, la Classification a pour but d'obtenir une représentation schématique simple d'un tableau rectangulaire de données dont les colonnes, suivant l'usage, sont des descripteurs de l'ensemble des observations, placées en lignes.

L'objectif le plus simple d'une classification est de répartir l'échantillon en groupes d'observations homogènes, chaque groupe étant bien différencié des autres. Le plus souvent,

cependant, cet objectif est plus raffiné ; on veut, en général, obtenir des sections à l'intérieur des groupes principaux, puis des subdivisions plus petites de ces sections, et ainsi de suite. En bref, on désire avoir une hiérarchie, c'est à dire une suite de partitions "emboîtées", de plus en plus fines, sur l'ensemble d'observations initial. [2]

2.3 Domaines d'application de la classification

La classification est en pratique appliquée dans la plupart des domaines du monde réel.

Nous la trouvons à titre d'exemple dans :

- Le Web, pour la classification des documents en fonction de leurs sujets et le filtrage des spam (spam/non spam) ;
- Le secteur médical, pour la classification des patients en fonction de leurs maladies ;
- La bio-informatique, pour la classification des gènes quand une grande quantité de gènes peuvent montrer des comportements similaires.
- Le marketing, pour la classification des entreprises en fonction de leurs productions. [3]

2.4 Les types de classification automatique :

La classification automatique consiste à regrouper divers objets (les individus) en sous-ensembles d'objets (les classes), on distingue deux approche de classification, la classification supervisée les classes sont connues à priori, contre la classification non-supervisée (en anglais clustering) les classes sont fondées sur la structure des objets.

2.4.1 La classification supervisée

Dans ce type de classification, les classes sont prédéfinies avec une description des documents. Lorsqu'un nouveau document arrive, on le compare avec la description de chaque classe et on le met dans celle qui lui ressemble le plus. Plusieurs techniques sont utilisées, on peut citer K voisins Proches, Arbre de Décision, Naïve Bayes, Machine à vecteur de support…

Dans ce qui suit notre travail va être concentré sur la catégorisation de textes (la classification supervisée)

2.4.2 La classification non supervisée

Les objets sont groupés dans des classes homogènes disjointes. Pour faire ressortir les ensembles de documents, on doit maximiser l'homogénéité interne des classes et la dispersion entre elles. Les deux méthodes principales du clustering sont : les méthodes hiérarchiques et les méthodes non hiérarchiques. [4]

3-Catégorisation de texte

3.1 Définition

La catégorisation automatique de textes est d'apprendre à une machine à classer un texte dans la bonne catégorie en se basant sur son contenu. La catégorisation automatique de textes est d'apprendre à une machine à classer un texte dans la bonne catégorie en se basant sur son contenu.

Dans une catégorisation de texte : la classification s'apparente au problème de l'extraction de la sémantique d'un texte, puisque l'appartenance d'un document à une catégorie est étroitement liée à la signification de ce texte.

3.2 Comment catégoriser un texte ?

Le processus de catégorisation intègre la construction d'un modèle de prédiction qui, en entrée, reçoit un texte et, en sortie, lui associe une ou plusieurs étiquettes. Pour identifier la catégorie ou la classe à laquelle un texte est associé, un ensemble d'étapes est habituellement suivies. Ces étapes concernent principalement la manière dont un texte est représenté, le choix de l'algorithme d'apprentissage à utiliser et comment évaluer les résultats obtenus pour garantir une bonne généralisation du modèle appris. Le processus de catégorisation, intégrant la phase de classement de nouveaux textes, est résumé dans la (figure 1.1). Il comporte deux phases que l'on peut distinguer comme suit :

3.2.1 L'apprentissage :

Qui comprend plusieurs étapes et aboutit à un modèle de prédiction :

a) nous disposons d'un ensemble de textes étiquetés (pour chaque texte nous connaissons sa catégorie) ;

b) à partir de ce corpus, nous extrayons les k descripteurs (mots, termes) (t1; ..; tk) les plus pertinents au sens du problème à résoudre ;

c) nous disposons alors d'un tableau « descripteurs × individus », et pour chaque texte nous connaissons la valeur de ses descripteurs et son étiquette ;

3.2.2 Le classement :

Le classement d'un nouveau texte dx, qui comprend deux étapes :

a) recherche puis pondération des occurrences (t1; ...; tk) des termes dans le texte dx à classer ;

b) application d'un algorithme d'apprentissage sur ces occurrences et le tableau précédent afin de prédire l'étiquette de ce texte dx. [5]

3.3 Processus de la catégorisation de texte :

Le processus reçoit en entrée un document textuel afin de lui trouver sa catégorie, pour cela plusieurs étapes doivent d'être suivies. Ces étapes sont :

- La représentation des textes
- La Pondération des termes
- La réduction de la taille du vocabulaire
- Choix de classificateur
- Evaluation du modèle

La figure 1.1 résume le processus de catégorisation des textes qui comporte deux phases : l'apprentissage et le classement. [6]

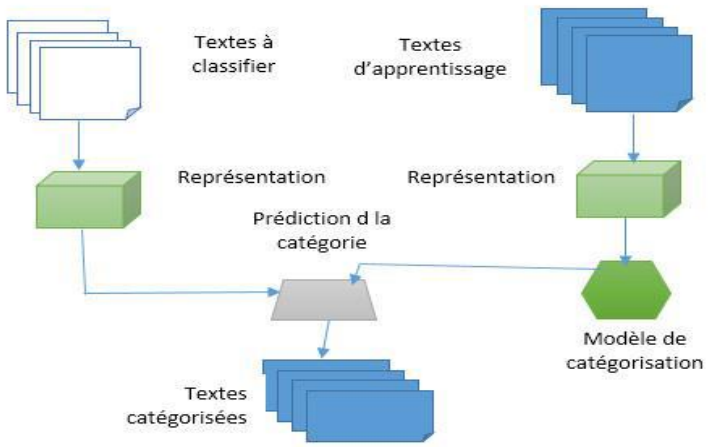

Figure 1.1 Processus de catégorisation de textes. [7]

3.4 Représentation de texte (choix des termes)

Dans la catégorisation de textes, comme dans la recherche documentaire, on transforme le document dj en un vecteur dj = (w1j, w2j, ..., w|T|j), où T est l'ensemble de termes (descripteurs) qui apparaissent au moins une fois dans le corpus (la collection) d'apprentissage. Le poids wkj correspond à la contribution du terme tk à la sémantique du texte dj. [5]

3.4.1 Représentation en sac de mots (bag of words)

Cette méthode consiste à représenter le document sous forme d'un vecteur de mots. Le processus qui permet de convertir le texte d'un document en un ensemble de termes est appelé l'analyse lexicale qui permet de reconnaître les espaces de séparation des mots, les ponctuations, les chiffres,…etc., pour qu'ils seront tous supprimés de la représentation. Cette représentation a comme avantage d'exclure toute analyse grammaticale et toute notion de distance entre les mots, mais présente comme inconvénient la difficulté de délimiter les mots dans certaines langues telles que l'Arabe ou l'Allemand.

3.4.2 Représentation avec les racines lexicales

Cette méthode consiste à remplacer les mots du document par leurs racines lexicales, qui peut être réalisée en utilisant un des algorithmes les plus connus pour la langue anglaise qui est l'algorithme de Porter [Porter, 1980] de normalisation de mots qui sert à supprimer les affixes de ces derniers pour obtenir une forme canonique. Cette méthode a comme avantage de regrouper les différentes flexions d'un mot dans une seule composante, et comme inconvénient la perte de sens car la racine extraite peut être commune à des mots se rapportant à des concepts différents.

3.4.3 Représentation avec les lemmes

Cette méthode consiste à remplacer les mots du document par leurs lemmes, elle doit utiliser l'analyse grammaticale afin de remplacer les verbes par leurs formes infinitives et les noms par leurs formes au singulier. En effet, Un mot donné peut avoir différentes formes dans un texte, mais leur sens reste le même.

3.4.4 Représentation avec les n-grammes

Cette méthode consiste à représenter le document par des n-grammes. Le n-gramme est une séquence de n caractères consécutifs. Cette technique présente plusieurs avantages. Les n-grammes capturent automatiquement les racines des mots les plus fréquents sans passer par l'étape de recherche des racines lexicales, indépendante de la langue, les espaces sont pris en considération parce qu'en effet, la non prise en compte de ces derniers introduit du bruit.

3.4.5 Représentation conceptuelle

Cette méthode consiste à représenter le document sous forme d'un ensemble de concepts, ces concepts peuvent être capturés en utilisant les réseaux sémantiques ou les sous arbres (un sous arbre représente une hiérarchie de concepts).

3.5 Pondération des termes (Codage des termes) :

La pondération des termes permet de mesurer l'importance d'un terme dans un document. Cette importance est souvent calculée à partir de considérations et interprétations statistiques (ou parfois linguistiques). Les méthodes les plus populaires sont :

3.5.1 Mesure TF (TermFrequency)

Cette mesure est proportionnelle à la fréquence du terme dans le document (pondération locale). Elle peut être utilisée telle quelle ou selon plusieurs déclinaisons (log (TF), présence/absence,. . .).

3.5.2 Mesure TFIDF (Term Frequency Inverse Document Frequency)

Le poids d'un terme T dans un document D est calculé comme suit :

$$TFIDF(T, D) = TF(T, D) * \log (N/ DF(T))$$

Avec :

TF(T, D) : la fréquence du terme dans le document,

N : le nombre total de documents de la base documentaire.

DF(T) : le nombre de documents contenant le terme. [57]

3.6 Domaines Applications de la catégorisation de texte :

La catégorisation de textes est utilisée dans de nombreuses applications. Parmi ces domaines figurent : l'identification de la langue, la reconnaissance d'écrivains, la catégorisation de documents multimédia, et bien d'autres.

La catégorisation de textes peut être un support pour différentes applications parmi lesquelles le filtrage, qui consiste à déterminer si un document est pertinent ou non (décision binaire), par exemple la détection de spams (les courriers indésirables) pour ensuite les supprimer, le routage qui permet d'affecter un document à une ou plusieurs catégories parmi n, par exemple la diffusion sélective d'information. Lors de la réception d'un document l'outil choisit à quelles personnes le faire parvenir en fonction de leurs centres d'intérêt. Ces centres d'intérêt correspondent à des profils individuels. [5]

3.7 Critères d'évaluation des classificateurs

Presque toujours, on divise le corpus de textes déjà classées et disponible en deux ensembles, l'ensemble d'entrainement sur lequel le classificateur fait son apprentissage et l'ensemble de test sur lequel on peut évaluer sa performance. Ils existent de nombreuses mesures pour calculer la performance d'un classificateur. Elles dépendent essentiellement du modèle de tâche pour lequel le système de catégorisation de texte (CT) est utilisé. Pour mieux illustrer les différentes mesures qui vont suivre, on prend pour point de départ la table de contingence illustré par le tableau ci-dessous, distincte pour chaque catégorie.

On définit à partir des statistiques de cette table les mesures suivantes :

Précision :

$$Précision = a/(a+b) \qquad (1)$$

Soit le nombre d'assignations correctes sur le nombre total d'assignations.

Rappel :

$$Rappel = a/(a+c) \qquad (2)$$

Soit le nombre d'assignations correctes sur le nombre d'assignation qui auraient dû être faites.

Exactitude :

$$Exactitude = (a+d)/(a+b+c+d) \qquad (3)$$

Erreur :

$$Erreur = (b+c) /(a+b+c+d) \quad (4)$$

Les deux dernières mesures, bien que couramment utilisées en apprentissage automatique, sont jugées moins adaptées à la tâche de classification de textes. La précision et le rappel sont les mesures les plus rencontrées dans la littérature. Lors de l'évaluation de la performance d'un classificateur, on ne peut tenir compte de la précision ou de rappel séparément.

Effectivement, on pourrait mettre en place un système qui rejetterait tous les textes: il obtiendrait une précision de 100%, mais un rappel de 0%. A l'inverse, un système qui accepterait tous les textes aurait un rappel de 100%, mais une précision de 0%. On voit donc qu'un meilleur classificateur est celui qui tente de faire le compromis idéal entre ces deux facteurs. Aussi, la mesure F1 est beaucoup utilisée. Elle est définie ainsi :

$$F1=2rp / (r+p) \quad (5)$$

Où : R représente le rappel et P représente la précision, F1est une fonction qui est maximisée quand la précision et le rappel sont proches. [7]

3.8 Les algorithmes d'apprentissage supervisé :

- ✓ **Machine à vecteurs de support :**

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais support vector machine, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Les SVM sont une généralisation des classifieurs linéaires. [8]

- ✓ **Réseaux de neurones**

Un réseau de neurones artificiels, ou réseau neuronal artificiel, est un ensemble d'algorithmes dont la conception est à l'origine très schématiquement inspirée du fonctionnement des neurones biologiques, et qui par la suite s'est rapproché des méthodes statistiques. [9]

✓ **Méthode des k plus proches voisins**

La méthode des k plus proches voisins est une méthode d'apprentissage supervisé. En abrégé k-NN ou KNN, de l'anglais k-nearest Neighbors. [10]

✓ **Arbre de décision**

Un arbre de décision est un outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteints en fonction de décisions prises à chaque étape. [11]

✓ **Classification naïve bayésienne**

La classification naïve bayésienne est un type de classification Bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Elle met en œuvre un classifieur bayésien naïf, ou classifieur naïf de Bayes, appartenant à la famille des classifieurs linéaires. [12]

3.9 Les problèmes de la catégorisation de texte :

Nous allons signaler les dix principales difficultés qui s'opposent à la catégorisation de textes :

➢ **Redondance(Synonymie)**

La redondance et la synonymie permettent d'exprimer le même concept par des expressions différentes, plusieurs façons d'exprimer la même chose. Cette difficulté est liée à la nature des documents traités exprimés en langage naturel contrairement aux données numériques.

➢ **Polysémie (Ambiguïté)**

A la différence des données numériques, les données textuelles sont sémantiquement riches, du fait qui sont conçues et raisonnées par la pensée humaine. Contrairement des langages informatiques, le langage naturel, autorise des violations des règles grammaticales engendrant plusieurs interprétations d'un même propos. Un mot possède, dans différents cas, plus d'un sens et plusieurs définitions lui sont associées.

- **L'homographie**

Deux mots sont dits homographes si 'ils s'écrivent de la même façon sans forcément avoir la même prononciation. L'homographie est une sorte d'ambigüité supplémentaire. (Ex : avocat en tant que fruit et avocat en tant que juriste) L'homographie et l'ambiguïté génère du bruit qui va causer une dégradation de précision (indicateur nécessaire pour mesurer la performance du classifieur). Il sera alors préférable d'ôter ces ambigüités.

- **La graphie**

Un terme peut comporter des fautes d'orthographe ou de frappe comme il peut s'écrire de plusieurs manières ou s'écrire avec une majuscule. Ce qui va peser sur la qualité des résultats. Parce que si un terme est orthographié de deux manières dans le même document (Ghelizane, Relizane), la simple recherche de ce terme avec une seule forme graphique néglige la présence du même terme sous d'autres graphies, ce qui va influencer les résultats puisque les différentes graphies vont être traitées séparément

- **Les variations morphologiques**

Les conjugaisons, pluriels, influent négativement sur la qualité des résultats puisque les différentes variations morphologiques vont êtes considérées séparément et chacune va être prise comme un élément à part comme par exemple les trois termes : maître, maîtresse, maîtriser sont traités indépendamment quoique en réalité ça pivote sur la même idée.

- **Les mots composés**

La non prise en charge des mots composés comme : comme Arc-en-ciel, peut-être, sauve- qui-peut, etc. Dont le nombre est très important dans toutes les langues, et traiter le mot Arc- en-ciel par exemple en étant 3 termes séparés réduit considérablement la performance d'un système de classification néanmoins l'utilisation de la technique des n-grammes pour le codage des textes atténue considérablement ce problème des mots composés.

- **Présence-Absence de termes**

La présence d'un mot dans le texte indique un propos que l'auteur a voulu exprimer, on a donc une relation d'implication entre le mot et le concept associé, quoique on sait très bien qu'il ya plusieurs façons d'exprimer les mêmes choses, dès lors l'absence d'un mot n'implique pas obligatoirement que le concept qui lui est associé est absent du document.

- **Complexité de l'algorithme d'apprentissage**

Nous avons représenté généralement sous forme de vecteur contenant les nombres d'apparitions des termes dans ce texte. Or, le nombre de textes qu'on va traiter est très important sans oublier le nombre de termes composant le même texte donc on peut bien imaginer la dimension du tableau (textes * termes) à traiter qui va compliquer considérablement la tâche de classification en diminuant la performance du système.

- **Sur-apprentissage**

Le nombre de termes très important et très varié qui ne se répètent dans tous les textes va causer énormément de creux dans le tableau de grande dimension (textes*termes) qui peut provoquer du sur-apprentissage qui s'explique par le fait que le modèle n'arrive pas à bien classer les nouveaux textes, pourtant il l'a bien fait dans la phase d'apprentissage en classant correctement les textes de la base d'apprentissage. Pour limiter le sur-apprentissage, on doit sélectionner des termes pour réduire la dimensionnalité. D'après les expériences antérieures, le nombre de termes doit être limité par rapport au nombre de textes de la base d'apprentissage.

- **Subjectivité de la décision**

Parmi les problèmes classiques usuels dans le domaine de l'apprentissage supervisé c'est la subjectivité de la décision prise par les experts qui décident de la classe à laquelle le texte va être attribué. [3]

4 Etat de l'art

La langue arabe est très riche en forme et grammaticalement. Il a consacré beaucoup d'efforts pour développer et servir à faciliter le processus de classification. Malgré cela, il y a encore un manque de classification des systèmes de textes arabes.

4.6 Les travaux de Laila Khreisat (2009)

Dans cette étude, le chercheur a montré approche automatisée pour la classification des textes en arabe. Il a utilisé deux méthodes de classification : tri-gramme où sont utilisées pour déterminer les bases des inclusions à cet effet (instance) en l'ensemble des nombres réels et la distance de Manhattan. Les résultats ont montré que la classification d'une échelle à tri-gramme que d'utiliser des amendes à l'aide Classé à l'échelle de Manhattan. [13]

4.7 Les travaux de Med El Amine Abderrahim (SD)

L'idée de cette étude est d'exploiter une ressource lexicale et un analyseur morphologique pour reformuler (par expansion) la requête de l'utilisateur afin d'améliorer les résultats de la recherche. Pour tester cette approche ils ont utilisé le moteur de recherche Google, pour envoyer La requête. Pour ensuite faire appel à WordNet Arabe pour enclencher le processus d'expansion. [14]

4.8 Les travaux de Karim Djelailia, AbdessalemKelaiaia, Hayat Farida Merouani (SD)

Cette recherche consiste à l'influence de la radicalisation et de la réduction de l'espace de représentation dans la qualité des résultats de classification de textes arabes. On a utilisé les machines à vecteurs supports (SVM) pour la classification avec le codage des termes TF-IDF. Ensuite une comparaison entre les résultats obtenus avec stemming et ceux utilisant directement les termes extraits à partir des textes du corpus. Les résultats sont très encourageants. [15]

4.9 Les travaux de EL KHADIR LAMRANI, EL HABIB BEN LAHMAR, ABDELAZIZ MARZAK (SD) :

Dans ce travail, ils ont utilisé trois algorithmes qui signifie k-means, k-means++ et classification hiérarchique ascendante pour de clustering des documents textes. Où chaque document appartient à un seul document. Enfin, ils ont évalué et comparé sur la base de plusieurs critères. Lorsque les résultats ont montré que chaque algorithme a ses limites et ses avantages. [16]

4.10 Les travaux deTarekKanan et Edward A. Fox (SD)

Dans cette étude, il a été procédé à une petite classification des histoires des besoins d'information arabes au Qatar et d'autres pays. Où il a été placé tige spécialement conçu et utilisé pour la classification l'algorithme Svm. Les résultats ont montré que l'approche adoptée mieux que les technologies modernes au côté de classification. [17]

5- Conclusion

La classification automatique des textes est l'un des éléments les plus importants du système de récupération de l'information, ce qui permet l'organisation des documents par catégorie et faciliter ainsi le processus de recherche, car une bonne classification exige une bonne représentation doit donc donner la représentation textuelle de l'importance des mêmes méthodes de classification.

CHAPITRE 2
LES ONTOLOGIES
CONCEPS ET NOTIONS DE BASE

1.Introduction :

Une ontologie en informatique est un ensemble structuré de concepts permettant de donner un sens aux informations. L'objectif premier d'une ontologie est de modéliser un ensemble de connaissances dans un domaine donné. Alors nous avons besoin de visualiser et de mettre à jour des informations dans ce domaine. Donc dans ce chapitre, nous présentons l'ontologie à tous égards.

2. Les ontologies

2.1 Présentation des ontologies :

Les ontologies sont employées dans l'intelligence artificielle, le Web sémantique, le génie logiciel, l'informatique ou encore l'architecture comme une forme de représentation de la connaissance au sujet d'un monde ou d'une certaine partie de ce monde. Les ontologies décrivent généralement :

— Individus : les objets de base ;

— Classes : ensembles, collections, ou types d'objets ;

— Attributs : propriétés, fonctionnalités, caractéristiques ou paramètres que les objets Peuvent posséder et partager ;

— Relations : les liens que les objets peuvent avoir entre eux ;

— Evénements : changements subis par des attributs ou des relations. [18]

2.2 Définitions :

***Définition 1** : une ontologie définit les termes et les relations de base du vocabulaire d'un domaine ainsi que les règles qui permettent de combiner les termes et les relations afin de pouvoir étendre le vocabulaire. [19]

*__Définition 2__ : une ontologie est une spécification explicite d'une conceptualisation. [20]

*__Définition 3__ : une ontologie est une théorie logique proposant une vue explicite et partielle d'une conceptualisation. [21]

*__Définition 4__ : une ontologie organise dans un réseau des concepts représentant un domaine. Son contenu et son degré de formalisation sont choisis en fonction d'une application. [22]

2.3 Les Composants d'une ontologie :

L'ontologie n'est en fin de compte qu'une modélisation du monde réel en concept et relation entre ces concepts. La formalisation d'une ontologie se met en place grâce à 5 types de composants. Les principales composantes qu'on peut distinguer sont donc les suivantes :

2.3.1 Les concepts

Concept ou classe, définissant un ensemble d'objet, abstrait ou concret, que l'on souhaite modéliser pour un domaine donné. Les connaissances portent sur des objets auxquels on se réfère à travers des concepts.

2.3.2 Les instances :

Instance ou individus, constituent la définition extensionnelle de l'ontologie (pour représenter les éléments spécifiques). [23]

2.3.3 Les fonctions

Fonctions constituent des cas particuliers des relations, dans laquelle un élément de la relation, le nième (extrant) est défini en fonction des n-1 éléments précédents (intrants). [18]

2.3.4 Les axiomes

Les axiomes sont utiles à la structuration de phrases qui sont toujours vraies. Ils permettent de contraindre les valeurs de classes ou d'instances.

2.3.5 Les relations

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. [24]

2.4 Les types d'ontologie :

On présente les types d'ontologies les plus couramment utilisés.

2.4.1 Les ontologies de haut-niveau (ontologies génériques) :

Elles sont similaires aux ontologies de domaine, mais les concepts qui y sont définis sont plus génériques et décrivent des connaissances telles que l'état, l'action, l'espace et les composants. Généralement, les concepts d'une ontologie de domaine sont des spécialisations des concepts d'une ontologie de haut niveau. [23]

2.4.2 Les ontologies spécialisées

Ce sont des ontologies qui « spécialisent » un sous-ensemble d'ontologies génériques en un domaine ou un sous-domaine. Elles peuvent être de domaine, d'application, techniques. [24]

Les trois principales sont :

*** Les ontologies de domaine :**

Définissent des conceptualisations spécifiques à certains domaines. Les méthodologies d'ingénierie des connaissances font une distinction explicite entre ontologies de domaine et connaissances du domaine : alors que les connaissances du domaine décrivent des situations effectives dans un certain domaine, l'ontologie de domaine pose les contraintes sur la structure et le contenu (i.e. la grammaire et le vocabulaire) des connaissances du domaine.

*** Les ontologies d'application :**

Ces ontologies décrivent des concepts dépendant à la fois d'un domaine et d'une tâche particulière, qui sont souvent des spécialisations des deux ontologies relatives. Ces concepts correspondent souvent aux rôles joués par des entités de domaine tout

en exécutant une certaine activité, comme l'unité remplaçable ou le composant disponible. [18]

*** Les ontologies de tâche :**

Ce type d'ontologie décrit un vocabulaire en relation avec une tâche ou une activité générique comme le diagnostic ou la vente. Les ontologies de tâche fournissent un lexique systématisé de termes utilisés pour résoudre les problèmes associés à des tâches particulières Elles incluent des noms génériques (par ex., plan, objectif, contrainte), des verbes génériques (par ex., assigner, classer, sélectionner), des adjectifs génériques (par ex., assigné) et d'autres mots qui relèvent de l'établissement d'échéances.

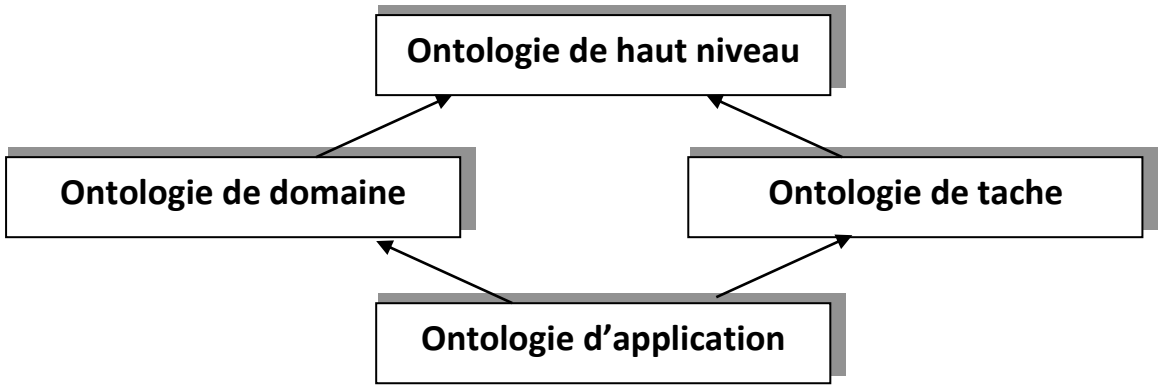

Figure 2.1:Type d'ontologie selon GUARINO

2.5 Construction d'une ontologie :

2.5.1 Principe de construction d'une ontologie :

Plusieurs travaux se sont intéressés à l'élaboration de principes de construction d'ontologies. Gomez a énuméré un certain nombre de principes à suivre pour l'élaboration d'une ontologie, inspirés par les différents travaux existants : Clarté et objectivité, Exhaustivité, Cohérence, Extensibilité, Interventions ontologiques minimales, Distinction ontologique, Minimisation des distances sémantiques entre les concepts frères. [24]

2.5.2 Cycle de vie d'une ontologie :

La construction d'ontologies s'inscrit dans un cycle de vie classique, composé de quatre phases nominales :

- **La spécification**

Permet de fixer le but de la construction de l'ontologie et les utilisateurs de celle-ci. Elle fixeles limite du domaine à modéliser et l'utilisation qui sera faite des connaissances qu'elle permet de représenter.

- **La conceptualisation**

Permet de structurer le domaine de connaissances à représenter. Il s'agit là de proposer un modèle identifiant et structurant les concepts du domaine d'étude.

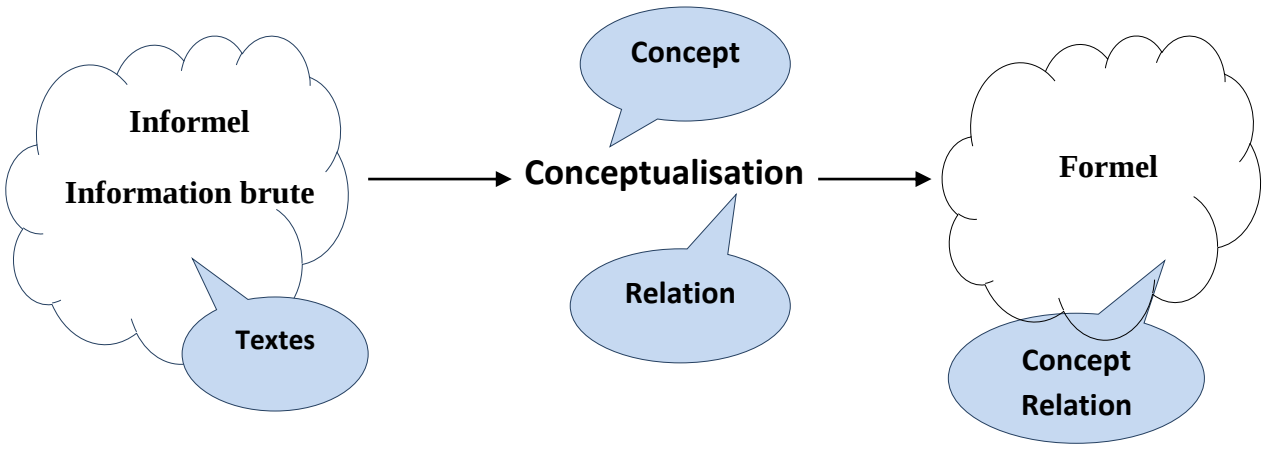

Figure 2.2: Conceptualisation d'une ontologie

- **La formalisation**

Permet le passage du modèle conceptuel obtenu dans la phase de conceptualisation à un modèle formel. Cette phase amène à choisir un langage de représentation des connaissances.

- **L'implémentation**

Permet de transformer le modèle formel en une entité manipulable par un système informatique. [18]

2-6. Classification des ontologies

Les ontologies peuvent être de nature très diverses. Afin de mieux s'y retrouver, un certain nombre de classifications ont été proposées.

La plus courante des classifications d'ontologies est la classification selon l'objet de conceptualisation [25]. On peut ainsi distinguer sept catégories [26] :

- **Les ontologies de représentation des connaissances** : les ontologies de représentations des connaissances sont utilisées pour formaliser un modèle de représentation des connaissances. On peut par exemple citer l'exemple de l'ontologie de frame [27], qui définit les primitives de représentation des langages à base de frames (classes, instances, slots, facettes, etc.).

- **les ontologies supérieures (aussi appelées ontologies de haut niveau):**
  Une ontologie de haut niveau décrit des concepts très généraux comme l'espace, le temps, la matière, les objets, les événements, les actions, etc. Ces concepts ne dépendent pas d'un problème ou d'un domaine particulier. Ces ontologies doivent être, du moins en théorie, consensuels à de grandes communautés d'utilisateurs [28]. Des exemples d'ontologies de haut niveau sont « UpperCyc ».

- **Les ontologies génériques (méta-ontologie)**:

  Elles contiennent des concepts généralistes, maismoins abstraits que ceux contenus dans les ontologies de haut niveau. On pourra réutiliser l'ontologiedans plusieurs domaines [25]. Un exemple d'une telle ontologie : SUMO (Suggested Upper Merged Ontology), une autre ontologie générique a été développée dans cette classe citons :Wordnet.

- **Les ontologies de tâches** [29] : Ce type d'ontologie sert à modéliser les tâches d'un problème ou d'une activité donnée. Ce type d'ontologie est utile pour décrire la structure d'une tâche de résolution de problème de manière indépendante du domaine concerné.
- **Les ontologies de domaine** : Elles sont réutilisables à l'intérieur d'un domaine donné et modélisent le vocabulaire à l'intérieur de ce domaine [26]. La plupart des ontologies existantes sont des ontologies de domaine [25].

- **Les ontologies de tâches-domaine** : ce sont des ontologies de tâches spécifiques à un certain domaine. Un exemple d'une telle ontologie est celui d'une ontologie des termes liés à la planification chirurgicale. [26]

- **Les ontologies d'application** :Il s'agit du type d'ontologie le plus spécifique [25]. Les concepts que l'on trouve dans ce genre d'ontologies modélisent les concepts d'un domaine particulier dans le cadre d'une application donnée.

2.7- Les langages de représentation et de manipulation d'ontologies

Plusieurs langages de représentation et de manipulation d'ontologies ont été développés. Dans cette section, nous faisons une rapide revue de ceux qui nous paraissent très représentatifs parmi les standards et recommandations du World Wide Web Consortium (W3C) 3 : RDF/RDFS, OWL, OWL 2, SKOS et SPARQL. [13]

2.7.1 RDF (Le Resource Description Framework) :

est un modèle de données pour les objets (ressources) et les relations entre eux fournissant des sémantiques simples pour ce modèle de données qui peuvent être représentés en XML. RDF permet de représenter des métadonnées à propos des ressources (identifiées par des URI) du web.

2.7.2 RDFS (Le Resource Description Framework Schéma) :

Est un vocabulaire de base pour décrire les déclarations RDF, au même titre que le XML-S pour le langage XML. Il ajoute à RDF la possibilité de définir des hiérarchies de classes et de définir les genres et les propriétés des ressources, d'assigner des contraintes spécifiques sur la nature des documents et de fournir des informations sur

l'interprétation des déclarations RDF. Les schémas RDF permettent donc de garantir qu'un document RDF est sémantiquement consistant.

2.7.3 OWL (Ontologies Web Language) :

Est un composant de l'activité Web Sémantique qui vise à rendre les ressources Web plus accessibles aux processus automatisés en ajoutant des informations qui décrivent le contenu Web. [30]

2.7.4 SKOS (Simple Knowledge Organization System) :

Recommandation du W3C depuis août 2009, SKOS9 est un modèle de données destiné à supporter des RTO, telles que les terminologies, les thésaurus ou encore les taxonomies. De plus, il permet de décrire très en détail le niveau lexical d'une ressource ; ce qui est particulièrement intéressant dans des contextes de modélisation multilingue. Il offre un mécanisme simple permettant de supporter la représentation de vocabulaires structurés.

2.7.5 SPARQL (SPARQL Protocol And RDF Query Language) :

Comme le langage SQL, SPARQL10 est un langage de requêtes destiné à interroger les bases de triplets RDF, appelées aussi triple stores. Il est devenu une recommandation du W3C depuis janvier 2008. Il utilise des patterns de graphe pour déterminer les triplets qui satisfont les conditions des requêtes. En principe, avec ce langage, on peut accéder à toute donnée du Web représentée au format RDF. SPARQL utilise une syntaxe inspirée de SQL et est à ce titre très similaire à ce langage. [30]

2.8 Les éditeurs d'ontologies :

On a distingué quelques éditeurs :

2.8.1 PROTEGE :

Est un éditeur open source très complet (beaucoup de plugins). Il est le plus célèbre de sa catégorie. On pourrait aussi dire de lui que c'est « modeleur d'ontologie. ». Il

propose deux types de modélisation d'ontologies : protégé-frames et protégé-owl. Les ontologies créées dans protégé peuvent être exportées dans les formats RDF(S), OWL et XML shema. [32]

2.8.2 OILED :

Est un éditeur qui a été développé par l'université de Manchester pour éditer des ontologies dans les langages de représentation OIL20. Les versions disponibles d'OilEd ne constituent pas un environnement complet pour le développement d'ontologies d'envergure.

2.8.3 JENA :

Est un environnement de travail open source en Java, pour la construction d'application web sémantique. JENA permet de manipuler des documents RDF, RDFS, OWL et SPARQL. Il fournit un moteur d'inférences permettant des raisonnements sur les ontologies. JENA est maintenant sous Apache Software Licence.

2.8.4 ONTOLIGA :

Fournit un environnement collaboratif distribué pour chercher, créer, éditer, modifier, et utiliser des ontologies en ligne. Le serveur peut supporter jusqu'à 150 utilisateurs actifs. Certains fournissent une description de leurs projets. L'environnement assiste l'utilisateur dans les taches de développement et le maintien et leur permet de partager leur ontologie avec d'autres utilisateurs.

2.8.5 DOE (Differential Ontologies Editor) :

A été développé à l'Institut National de l'Audiovisuel par R. Troncy et A. Isaac en 2002. L'éditeur DOE offre des interfaces de création, modification et suppression de concepts et de relations, une représentation graphique de l'arbre ontologique, et des fonctionnalités de recherche et de navigation dans la structure créée. L'ontologie est documentée par des définitions encyclopédiques avec des synonymes et les principes différentiels en plusieurs langues. [32]

2.9 Utilisation des ontologies :

Même si le besoin de développer une ontologie est très varié et dépend du domaine d'application, nous pouvons facilement énumérer un certain nombre d'utilités, notamment :

• La connaissance du domaine : Les ontologies permettent la modélisation des connaissances dans un domaine particulier, dans lequel opère le système à développer.

• La communication : les ontologies assurent une communication fiable et hétérogène entre personnes et machines (agents logiciels ou organisations) du fait qu'elle permet de mettre en place un langage ou un vocabulaire conceptuel commun.

• L'interopérabilité : La représentation explicite des connaissances dans un domaine donné sous forme d'une ontologie, permet à son tour une plus grande réutilisation, un partage plus large et une interopérabilité plus étendue.

• L'aide à la spécification des systèmes : La représentation conceptuelle des éléments du domaine, permet aux systèmes de réaliser des raisonnements logiques qu'on appelle inférences, et de sortir avec des conclusions capables d'aider l'utilisateur ou le gestionnaire dans ses décisions.

• L'indexation et la recherche d'information : Dans le web sémantique, d'une façon générale, et dans notre application en particulier, les ontologies sont utilisées pour indexer et décrire les ressources utilisées. Cela permet une plus grande précision dans les résultats des recherches ou d'assignation des ressources. [23]

2.10WordNet ET ArabWordNet(AWN) :

2.10.1 WordNet:

WordNet est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton depuis une vingtaine d'années. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Des versions de WordNet pour d'autres langues existent, mais la version anglaise est cependant la plus complète à ce jour. La base de données ainsi que des outils sont disponibles gratuitement. Par rapport aux outils fournis, un développeur peut aussi accéder à la

base de données à partir des interfaces disponibles pour de nombreux langages de programmation.

WordNet est distribué sous une licence libre, permettant de l'utiliser commercialement ou à des fins de recherche. Chapitre 2 - Les ontologies

La dernière version distribuée en avril 2013 est la 3.1. Cette version est par ailleurs consultable en ligne. [32]

2.10.2 Arabe WordNet (AWN) :

AWN est une ressource lexicale de la langue Arabe disponible gratuitement. Elle est basée sur la conception et le contenu de Princeton WordNet et peut être liée à d'autres WN pour d'autres langues, ce qui permet une traduction de et à des dizaines de langues. En plus, la connexion de WordNet à l'ontologie SUMO (Suggested Upper Merged Ontology). Par ailleurs, la plateforme Amine prend en charge des traitements intelligents et la possibilité de définition de règles d'inférence donnant lieu ainsi à des opportunités d'utilisation de l'aspect sémantique dans les applications du Traitement Automatique de la Langue Arabe (TALA).

Arabe WordNet a les quatre composantes (balises) :

*Objet : les concepts des termes.

*Mot : les termes (mots).

*Forme : la racine des termes.

*Lien : les relations entre les concepts. [34]

3. Conclusion :

Dans ce chapitre, nous avons vu une présentation et une définition d'ontologies ; en définissant les composants d'ontologies ses types et le principe de construction. Comme nous avons parlé de certains éditeurs et des langages de représentation d'ontologies et de ses utilisations. Des recherches sont actuellement en cours dans ce domaine par référence aux systèmes de recherche d'informations. Une description des langages utilisés pour les manipuler.

CHAPITRE 3
L'ONTOLOGIE LEXICALE WORDNET

1. Historique et origine

Machiavel a dit un jour en politique, « la fin justifie les moyens ». L'absence d'un dictionnaire électronique facilement accessible, a fait de WordNet un projet d'étude (développement manuel) au début des années 1980 à l'Université Princeton par une équipe de psycholinguistes et de linguistes, qui se sont basées sur la mémoire lexicale humaine.Peu temps après, il a émergé pour devenir une ressources lexicale électronique de la langue anglaise, comprenant plus de 200.000 de mots de classe ouvertes ainsi que plus de 115.000 ensemble de synonymes. [36] Actuellement plusieurs réalisations descendantes de WordNet existent (différentes langues), parmi eux EuroWordNet (EWN, 1996) et ArabicWordNet (AWN), ce dernier est construit selon les méthodes développées pour EuroWordNet.

Deux éléments ont participé au succès de WordNet :

- La maturité du projet rendue possible grâce à un travail de plus de dix ans.

- Le libre accès aux sources du projet tant pour consultation que pour la modification ainsi que la possibilité de redistribution du produit modifié.

2. Présentation de WordNet

Qu'est-ce WordNet ? Un dictionnaire ? Un thésaurus ? Les dictionnaires contiennent généralement des connaissances sur des lexiesalors que les encyclopédies contiennent des connaissances éparses, du monde, sur la surface de la terre. Quant aux thésaurus, leur structure est bâti autour des concepts et aident l'utilisateur à acquérir l'unité lexicale la plus appropriée lorsqu'il a un concept à rechercher. WordNet n'est ni un dictionnaire classique ni un thésaurus : il est en fait, un arrangement des traits de chacune de ces deux ressources lexicales. [37]

Dictionnaire : Recueil des mots d'une langue, des termes d'une science, d'un art, rangés par ordre alphabétique, avec leur signification.

Thésaurus : une liste de termes sur un domaine de connaissances, reliés entre eux par des relations synonymiques, hiérarchiques et associatives. Le thésaurus constitue un vocabulaire normalisé.

Lexie : C'est une suite de caractères formant une unité sémantique, un mot, et pouvant constituer une entrée de dictionnaire.

Une encyclopédie peut prendre la forme d'un livre ou plusieurs livres. Elle se présente souvent comme une collection d'articles traitant chacun un thème.

Entités nommées : Sa reconnaissance est une sous-tâche de l'activité d'extraction d'information dans des corpus documentaires. Elle consiste à rechercher des objets textuels (c'est-à-dire un mot, ou un groupe de mot) catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc.

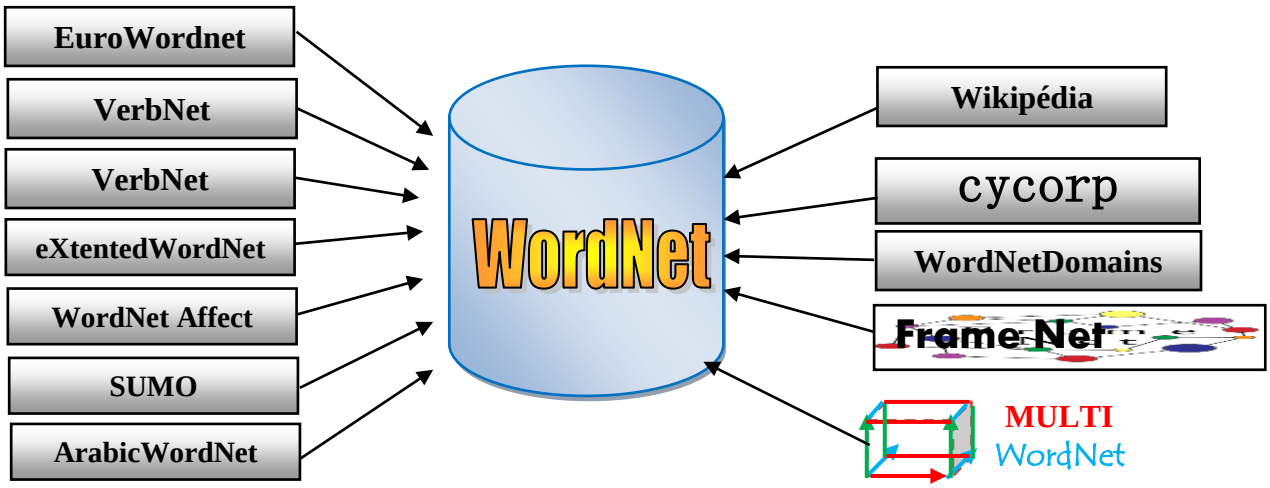

Figure 3.1 : Ressources descendances de WordNet
(Liste non exhaustive)

3. Conception & Structure de WordNet

On peut considérer WordNet comme un graphe ou un réseau sémantique, souvent qu'on qualifie d'ontologie légère (Light Ontology), ou chaque nœud représente un concept du monde réel. La conception de WordNet est basée sur les théories de la représentation des connaissances mentales : mémorisation des mots et concepts d'une manière hiérarchique, en utilisant la relation d'inclusion (qui lie, par exemple, des triplets comme «animal », « oiseau», et « Chardonnet »). [38]

3.1. SynSet

WordNet manipule les unités lexicales non pas par des mots mais par un ensemble de synonymes ou « Synset », groupes de mots ou de phrases qui expriment le même concept. Des différences de sens entre les membres d'un «Synset» se montrent dans différentes restrictions de sélection. Par exemple, « rise » (monter) et « fall » (tomber / descendre) peuvent choisir comme argument des entités abstraites comme « température » (température) et « prices » (prix). Un « Synset » est accompagné d'une petite définition dite « gloss » qui décrit un concept du monde réel.

Notons que WordNet met l'accent sur les liaisons entre les « Synset » (arc du graphe de la Figure 3.2) pour marquer sa valeur ajoutée faces aux dictionnaires traditionnaires. Chaque lien décrit une relation entre concept du monde réel. Par example, les relations tellesque : « a spoke is a part of a wheel » ou « a vehicle is a kind of conveyance »

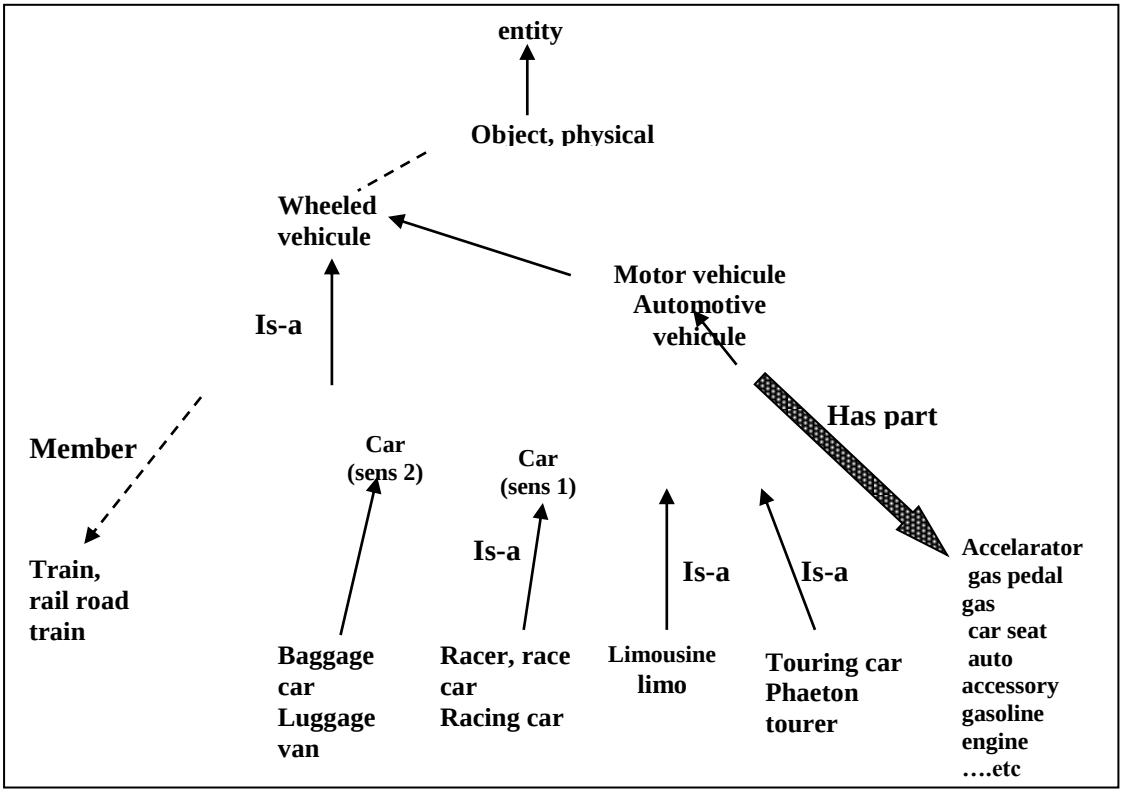

Figure. 3.2: Exemple de sous hiérarchie dans WordNet correspondent au concept "car". [39]

3.2. Organisation

WordNet sépare les données en quatre (04) bases de données, organisées différemment les unes des autres, associées aux catégories de noms, verbes, adjectifs et adverbes. Les noms et verbes sont organisés en hiérarchies. Des relations d'hyperonymie («est-un») et d'hyponymie relient les « ancêtres » des noms et des verbes avec leurs «spécialisations ».[FEL3798]

Word Meaning	Word Forms			
	F1	F2	F3	Fn
M1	E_{11}	E_{12}		
M2		E_{22}		
M3			E_{33}	
….			….	
Mn	E_{mn}			

Tableau 3.1 : Illustration des concepts de la matrice lexicale [37]

4. Les relations dans WordNet

Deux relations fondamentales interviennent dans WordNet, notamment celle entre les« wordform » appelés relations lexicales (par exemple : la synonymie), et celle qui associent les «wordmeaning » appelés relation sémantiques (par exemple : l'hyperonymie).

Remarquons que la majorité des relations dans WordNet sont des « synset » de la même catégorie, excepté les relations « pertains to » et « attribute » souvent utilisées entre les adjectifs et les noms. Le tableau 2 illustre un sous ensemble des relations dans WordNet qu'on détaillera par la suite.

Relation	Description	example
Hypernym	Is a generalization of	Furniture is a hypernym of chai
Hyponym	Is a kind of	Chair is a hyponym of furniture
Troponym	Is a way to	Amble is a troponym of walk
Meronym	Is part/substance/member of	Wheel is a (part) meronym of a bicycle
Holonym	Contains part	Bicycle is a holonym of a wheel
Antonym	Opposite of	Ascend is an opposite of descend
Attribute	Attribute of	Heavy is a attribute of weight
Entailment	entails	To offend causes to resent
Cause	Cause to	To lodge is related to reside
Alsosee	Relatedverb	Dead is similar to assassinated
Similar to	Similar to	Dead is similar to assassinated
Participle of	Is participle of	Stored (adj) is the participle of "to store"
Pertainym of	Pertains to	Radial pertains to radius

Tableau 3.2 : Quelques relations dans WordNet

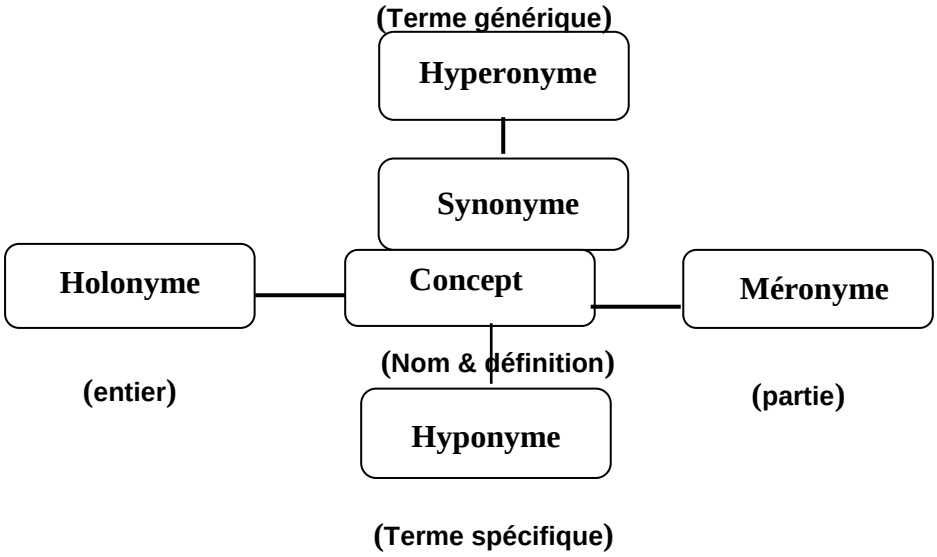

Fig. 3-3 : Principales relations sémantiques dans WordNet. [39]

Voilà, une définition des relations les plus importantes dans WordNet :

4.1. Synonymie

La synonymie est une relation liant deux concepts équivalents ou voisins (frêle / fragile). Il s'agit d'une relation symétrique.

4.2. Antonymie

Une autre relation familière est l'antonymie, qui s'avère être étonnamment difficile de définir. L'antonyme d'un mot « x » n'est pas toujours « Non x ». Par exemple, « riche » et « pauvre » sont des antonymes, mais de dire que quelqu'un n'est pas riche ne signifie pas qu'il doit être pauvre. [37]

4.3. L'Hyperonymie / Hyponymie

Elles sont réservées seulement pour les catégories Nom et Verbes qui se voient organisées sous forme d'une hiérarchie comportant un seul nœud racine. Les nœuds représentant les concepts les plus généraux sont les ancêtres des nœuds représentant les concepts les plus spécifiques.

Exemple :

dans la figure 3-1, « Entity » est le concept le plus hiérarchique des noms c'est la racine du nœud, il subsume le concept spécifique « wheeledvehicle » qui à son tour subsume le concept « Car ».

4.4. Méronymie

C'est une relation liant un concept C1 à un concept C2 qui est en fait une partie de C1 (C1= Fleur / C2= Pétale), un de ses membres (Forêt / Arbre) ou une substance le constituant (vitre / verre). Donc la méronymie est interprétée de trois manières différentes :

Il existe d'autres relations que la relations que nous allons donner juste une bref définition, telles que :

• Métonymie (Holonymie) : relation liant un concept C1 à un concept C2 dont il est une des parties. C'est la relation inverse de la méronymie.

• Implication : relation lie un concept C1 à un concept C2 qui en découle (marcher/faire un pas).

• Causalité : relation liant un concept C1 à son effet (tuer / mourir).

• Valeur : relation liant un concept C1 (adjectif) qui est un état possible pour un concept C2 (pauvre / condition financière).

• A pour valeur : relation liant un concept C à ses valeurs (adjectifs) possibles (taille / grand). C'est la relation inverse de Valeur.

• Voir aussi : relation entre des concepts ayant une certaine affinité (froid / gelé).

• Similaire à : certains concepts adjectifs dont le sens est proche sont regroupés. Un Synset est alors désigné comme étant central au regroupement. La relation « Similaire à » lie un Synset périphérique au Synset central (moite / humide).

• Dérivé de : indique une dérivation morphologique entre le concept cible (adjectif) et le concept origine (froidement / froid).

WordNet contient approximativement 117798 mots de nom organisés approximativement en 82115 concepts (Synset) (tableau 1) jusqu'à juillet 2008. Puisque la majorité des noms communs parfois sont des noms propres, aucune tentative curieuse de lesexclure n'est faite. En termes d'exhaustivité le but de WordNet diffère un peu des dictionnaires standards des écoliers. C'est à l'organisation de ces informations que WordNet espère de l'innovation.

Réseaux	**Formes**	**Synsets**	**Paires mots - sens**
Noms	117798	82115	146312
Verbes	11529	13767	25047
Adjectifs	21479	18156	30002
Adverbes	4481	3621	5580
TOTAL	155287	117659	206941

Tableau 3.3. Statistique sur WordNet (juillet 2008)

5. Les verbes dans WordNet (réseau sémantique)

Actuellement, WordNet contient plus de 25000 mots verbes. Les verbes sont divisés en 15 fichiers, selon un critère sémantique, presque chacun représente ce que les linguistes appellent domaine sémantique :

Les verbes de soins de corps et fonctions, changement, cognition, communication, compétition, consommation, contact, création, émotion, mouvement, perception, possession, interaction sociale et les verbes météorologiques. Pratiquement tous les verbes dans ces fichiers dénotent des évènements et des actions, un autre fichier contient les verbes d'état, tel que suffice, belong, et ressemble, qui ne peuvent pas être intégrés dans les autres fichiers. Les verbes dans ce dernier groupe font référence à un état, et ne constituent pas un domaine sémantique et ne partage aucune propriété sémantique.

6. L'hyponymie entre les verbes

La phrase modèle utilisée pour tester l'hyponymie entre les noms, « x est-un y » n'est pas convenable pour les verbes : to amble is a kind of to walk ce n'est pas une phrase correcte.

La distinction sémantique entre les verbes est différente des propriétés qui distinguent deux noms dans une relation hyponymique.

7. Polysémie1

Bien que les phrases anglaises nécessitent des verbes et non pas nécessairement des noms, le langage a moins de verbes que de noms. Les verbes sont polysémiques beaucoup plus que les noms.

8. ArabicWordNet (AWN)

8.1. L'écriture arabe [40]

L'arabe est une langue sémitique. Le système d'écriture de la langue arabe a 25 consonnes et trois voyelles longues : « و ،ا ،ي » : (OU, A, iii) on les appelle «**حروف العلة**» qui sont écrites de droite à gauche et prennent différentes formes en fonction de leur position dans le mot. En plus des voyelles longues, l'arabe a des voyelles courtes qui ne font pas partie de l'alphabet, mais plutôt sont écrites comme des voyelles diacritiques en

haut ou en bas d'une consonne pour lui donner le son désiré et par conséquent de générer un mot dans un sens souhaité.

Le terme « arabe classique » renvoie la forme standard de la langue utilisée dans tous les écrits et entendu à la télévision, la radio et dans les discours publics et les serments religieux. Les textes sans voyelles sont considérés comme étant plus appropriée par la communauté de langue arabe puisque c'est la forme habituelle de la vie quotidienne des documents écrits et imprimés (livres, magazines, journaux, lettres, etc.) Mais quand il s'agit du texte du « Coran », et plus généralement aux collections imprimées des livres scolaires et des dictionnaires arabes sur les supports en papier, les voyelles diacritiques apparaissent dans leurs intégralités. Comme on remarque très souvent dans des livres bien édités, des manuscrits ou bien certains textes imprimés la présence partielle ou au hasard des voyelles diacritiques sur les mots ambigus ou difficiles à lire.

Par exemple, un mot en arabe composé de deux lettres comme « **بر** » c'est à dire, «b» et «r», peut être très ambiguë, sans les voyelles diacritiques (voir Tableau 4), ou par exemple «**علم**». Notamment pour un écrivain, il peut utiliser les signes diacritiques afin que les lecteurs puissent facilement résoudre toute ambiguïté.

Arabic Word	Transliteration	POS	Meaning
بر	barr	noun	Land
بر	barr	adj	reverent
بر	burr	noun	wheat
بر	birr	noun	Reverence kindness

arabe	Translitération	POS	Sens
علم	‘alam	n	flag
علم	ilm	n	science
علم	ulima	v	known
علم	‘allama	v	teach
علم	‘alam	a	famous

Tableau 3.4 : Voyelles diacritiques possibles sur «بر » et sur « علم »

Pourtant, une mauvaise utilisation d'un seul signe diacritique fera un échec lors d'une requête, de recherche d'un document numériquepar exemple, « **السكون**» qui indique que la consonne n'est pas suivi par une voyelle, ou la « **الشدة** » (comme dans « **بر** » 'barr' dans le tableau 3-4 et « **درسّ** » darrasa dans le tableau 5),ce qui indique une double consonne {le premier n'est pas suivie d'une voyelle (ici «**السكون**») et la seconde est suivie d'une voyelle}.

Arabic Word	POS	Patterns	Meaning
Darasa درس	Verb	فعلfa?ala	Study
Darrasa درس	verb	فعل fa??ala	Teach
Dars درس	noun	فعلfa?l	lesson
Dira :sah دراسة	noun	فعالةfi?a:lah	Study
Mudarris مدرس	noun	مفعلmufa??il	teacher
Madrasah مدرسة	noun	مفعلةmaf?ala	school
Tadrisتدريس	noun	تفعيلtaf?i:l	teaching
Tada :rasa تدارس	verb	تفاعل tafa:?ala	discuss
Dirasi :دراسي	adj	فعاليfi?a:li	educational

Tableau 3.5 : Dérivations de la racine (d r s)

Beaucoup de personnes ont tendance à faire des erreurs sur la position de certains signes diacritiques sur un mot. Cela peut poser un sérieux problème pour les systèmes de recherche d'information et les systèmes informatisés de ressources lexicales qui dépendent de l'entrée, bien formulée, de l'utilisateur. Sinon, on peut assister même à des rejets de requêtes d'utilisateurs.

En particulier, il peut y avoir un rejet total d'une nouvelle ressource lexicale solide tel qu'AWN à moins que cette nouvelle ressource suppose que la plupart des utilisateurs du discours arabe ne sont pas experts en écriture des voyelles diacritiques et par conséquent les ignorent complètement. Ces utilisateurs sont plus à l'aise quand à la lecture de textes sans signes diacritiques, dans les documents écrits de tous les jours y compris les contrats juridiques et commerciaux, journaux, livres ainsi que les dictionnaires sur des supports papiers ou numérisés. En conclusion, on peut dire qu'il est préférable de permettre aux utilisateurs d'entrer des mots en arabe sans voyelles diacritiques mais parallèlement permettre au système de retrouver ces mots avec des voyelles diacritiques pour les besoins de désambiguïsation. [40]

8.2. Description d'AWN

L'ArabicWordNet est une base de données lexicale. Sa conception basé sur Princeton WordNet est construite suivant des méthodes développées pour EuroWordNetest

reliée avec l'ontologie SUMO (Suggested Upper Merged Ontology). ArabicWordNet a été développé par DOI / REFLEX (2005-2007) [40] (voir Figure 3.4).

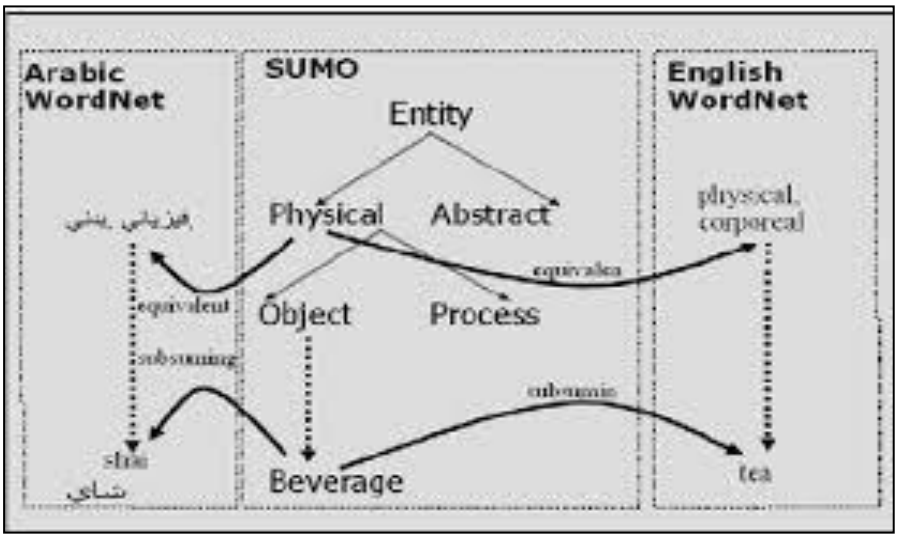

Fig. 3-4: Mapping de SUMO versWordNet(s)
(Structure et organisation de l'AWN)

SUMO est en pleine expansion afin d'offrir un fondement solide pour la formalisation sémantique de l'ArabicWordNet (AWN). La base de données AWN est disponible gratuitement. L'ontologie ArabicWordNet contient 9228 concepts « Synsets » (6252 nominales et 2260 verbales, 606 adjectival, et 106 adverbales), contient 18,957 expressions et 1155 concepts nommés [40] le fichier base de l'AWN sous format XML contient les quatre balises :

- Item : Contient les concepts (Synsets), les classes et les instances de l'ontologie.
- Word : Contient les mots arabes vocalisés.
- Form : Contient les Racines des mots arabes « root ».
- Link : Contient les relations entre les concepts.

8.3. Construction d'ArabicWordNet (AWN)

AWN sélectionne les Synsets, en se basant sur des critères [40] :

AWN doit être aussi dense que possible par des connexions de chaînes hyperonymie/hyponymie etc. En contreparties, La plupart des Synsets d'AWN doivent

correspondre à leurs homologues du WordNet anglais et la topologie entière des deux Wordnets doit être similaire.

Pertinence :

Donner la priorité aux concepts les plus fréquents. Ces critères incluront la fréquence des éléments lexicaux (en arabe et en anglais), PoS (Nom, verbes, adjectifs, adverbes…) et la fréquence des racines arabes dans leur corpus de référence respectifs.

Généralités :

Les Synsets les plus préférés sont ceux des ontologies de hauts niveaux – WordNet. Pour assurer ces trois critères, deux façons de procéder :

• **De l'anglais vers l'arabe** : On sélectionne, pour chaque Synset anglais, toutes les variantes correspondantes en arabe.

• **De l'arabe à l'anglais** : Tous les sens d'un mot arabe doivent être trouvés dans le WordNet anglais, ainsi que chacun de ces sens il faut sélectionner ces Synsets correspondants en anglais.

Ces deux étapes doivent être suivies tout au long de la construction d'AWN. Tous les Synsets AWN doivent être validé manuellement (et éventuellement verrouillé, lorsque toutes leurs variantes ont été trouvées), mais il convient d'exploiter, autant que possible, les ressources disponibles pour guider le processus de construction et de validation.
Une fois qu'un nouveau verbe arabe est ajouté à AWN, plusieurs possibilités d'extension sont à considérer : Les extensions des entrées verbales, y compris les dérivés verbaux1, les nominalisations et les noms verbaux, etc. Nous considérons également les formes les plus productives comme les dérivés des pluriels brisés(**جموع التكسير**). Ceux-ci peut-être fait grâce à un ensemble de règles lexicales et morphologiques pour tirer un maximum de profit de ces extensions des itérations courtes seront effectuées.

Pour construire AWN, il faut d'abord construire l'ensemble de la base de ses concepts (C.B.) à partir de l'ensemble de la base commune des concepts (CBCs4) d'EWN (EuroWordNet est un projet visant à construire des ontologies similaires au projet WordNet de l'université de Princeton pour 8 langues européennes dont le français) et BalkaNet (BalkaNet est un projet européen 2001-2004, visant à développer des Wordnets alignés pour la région des langues Balkans suivantes : Bulgare, Grec, Romanie, Serbe,

Turc et d'étendre le WordNet Tchèque précédemment élaborée dans le projet EuroWordNet) .

La concentration est faite sur les termes les plus pertinents afin d'obtenir environ 1.000 Synsets nominal et 500 Synsets verbale.

La préparation :

La préparation consiste au traitement des ressources disponibles bilingue et la compilation d'un ensemble de règles lexicales et morphologiques. De l'ensemble des dictionnaires bilingues disponibles, un dictionnaire bilingue homogène (HBIL) a été construit comprenant pour chaque information en entrée Arabe/Anglais, une paire de mot, la racine arabe est ajoutée manuellement, PoS (Part of speech), les fréquences relatives et les sources supportent l'appariement.

Dix sept (17) méthodes heuristiques sont utilisées pour le développement d'EWN et sont appliquées à HBIL [41] pour dériver les mots candidats anglais/arabe par un simple mappage des Synsets. Pour chaque mappage, l'information attachée comprend le mot arabe et sa racine, le Synset anglais, POS, les fréquences relatives, l'évaluation du mappage, la profondeur absolue dans WordNet, un certain nombre d'écarts entre le Synset et le sommet de la hiérarchie WordNet et les sources contenants la paire.

Les mots arabes dans les ressources bilingues doivent être normalisée et lemmatisée [42], [43], mais les voyelles et les signes diacritiques doivent être maintenus. Les racines arabes n'ont pas de voyelles.

Extension :

Après le prétraitement, l'ensemble des mots marqués arabe/anglais des paires Synset deviennent une entrée à l'étape de validation manuelle. Nous procéderons par blocs d'unités connexes (ensembles de SynsetsWordNet connexes, par exemple les chaînes d'hyponymie et l'ensemble des mots arabes connexes (C'est à dire, les mots ayant la même racine) au lieu des unités individuelles (Synsets, sens, mots). [40]

Finalement, AWN sera complété par l'ajout d'une terminologie et des entités nommé [44], pour combler les lacunes de sa structure qui couvre certain domaine spécifique.

8.4. L'interface Utilisateur

Outre la recherche et la navigation simple et facile sur l'ensemble de la base de données pour les utilisateurs finaux, les lexicographes ont besoin aussi d'une interface d'édition.

Une variété de composants hérités sont disponibles, chacun d'eux avec ses avantages relatifs.

Il s'agit de permettre aux multiples lexicographes de différents sites de maintenir une base de données commune.

9. Conclusion

WordNet est sans doute le précurseur et la référence en matière de base lexicales sémantiques et informatiques est devenue peu à peu à caractère ontologique, une description sommaire de ce dernier à était faite et constitue donc un exemple concret de notre présentation des ontologies. Nous avons tenté à travers ce chapitre, de donner un aperçu global de ce qu'est l'ontologie WordNet. Cette ressource lexicale nous a surpris tant dans sa structure que dans le travail et les efforts investis pour la réaliser. Ce chapitre est un avant plan de ce qu'on prévoit à étudier comme approche à la réalisation d'un WordNet arabe. Le chapitre suivant donne un état de l'art des différentes méthodes de l'apprentissage ontologique.

CHAPITRE4
IMPLEMENTATION ET EXPERIMENTATION

1. Introduction

Pour avoir une meilleure solutiona notre problématique qui est l'extraction des donnes a partir de l'ontologie lexicale WORDNET commebut principal a ce projet on a développé une application qui permette la classification supervisée des textes pilotée par l'ontologie WORDNET.

En effet, pour réaliser cette application on a utilisé des outils qui nous ont aidé au longdes différentesétapes de création de cette application, parmi ces outils on peut citer les suivants :

2. Description des outils

Cette application offre de multiples fonctionnalités qui exigent l'utilisation de différents outils pour les implémenter, certains de ces outils sont communs sur tous les projets java mais d'autres sont particuliers à ce type de projets.

2.1 Java

Java est un langage de programmation crée par Sun Microsystems en 1995. Beaucoup d'applications et de sites Web ne fonctionnent pas si Java n'est pas installé, et leur nombre ne cesse de croître chaque jour. Java est rapide, sécurisé et fiable. Des ordinateurs portables aux centres de données, des consoles de jeux aux superordinateurs scientifiques, des téléphones portables à Internet, la technologie Java est présente sur tous les fronts.

En a choisi java avec sa version 8 comme langage de programmation du projet.

2.2 NetBeans

NetBeans est un environnement de développement intégré (EDI), placé en open source par Sun en juin 2000 sous licence CDDL (Common Development and Distribution License) et GPLv2. En plus de Java, NetBeans permet la prise en charge native de divers langages tels le C, le C++, le JavaScript, le XML, le Groovy, le PHP et le HTML, ou d'autres (dont Python et

Ruby) par l'ajout de greffons. Il offre toutes les facilités d'un IDE moderne (éditeur en couleurs, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).

Compilé en Java, NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X ou sous une version indépendante des systèmes d'exploitation (requérant une machine virtuelle Java). Un environnement Java Development Kit JDK est requis pour les développements en Java.

NetBeans constitue par ailleurs une plate forme qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)). L'IDENetBeans s'appuie sur cette plate forme.

L'IDE Netbeans s'enrichit à l'aide de greffons.

NetBeans est aussi une plate-forme générique pour le développement d'applications pour stations de travail (bibliothèque Swing (Java)).Elle fournit des ressources pour développer les éléments structurants de ces applications: gestion des menus, des fenêtres, configuration, gestion des fichiers, gestion des mises à jour… Des présentations détaillées sont fournies par le centre de documentation de NetBeans.

L'IDE NetBeans comprend toutes les ressources utiles mais il est aussi possible d'installer la plate-forme séparément.

Le développement d'applications sur la base de la plate-forme NetBeans consiste en la réalisation de « modules » qui s'insèrent dans la plate-forme et en étendent dynamiquement les fonctions.

Un module est un groupe de classes Java, de portée variée : elle peut consister en une simple classe Java réalisant des fonctions simples (exemple : ajouter une action dans un menu pour éditer le contenu du presse papier) comme elle peut intégrer une application externe complète (exemple : Java profiling suite). Un module peut s'appliquer à l'IDE NetBeans lui-même.

La réalisation des modules s'appuie sur une API normalisée.

Un espace de partage entre développeurs est mis en place.

2.3 WORD NET :

WordNet est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton depuis une vingtaine d'années. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Des versions de WordNet pour d'autres langues existent, mais la version anglaise est cependant la plus complète à ce jour.

Figure 4.1 : NetBeans, Java, Wordnet

3. Interfaces Graphiques de l' application :

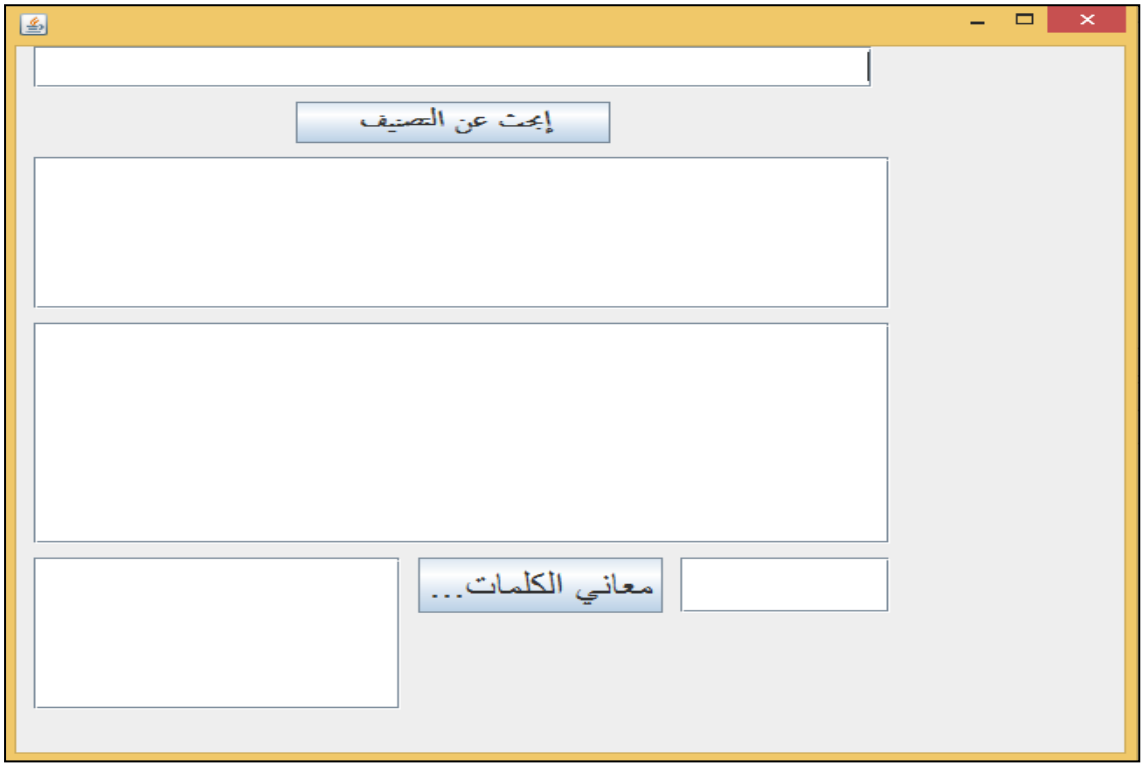

Figure 4.2Interface Principale

- La saisie du terme ou texte a classifié

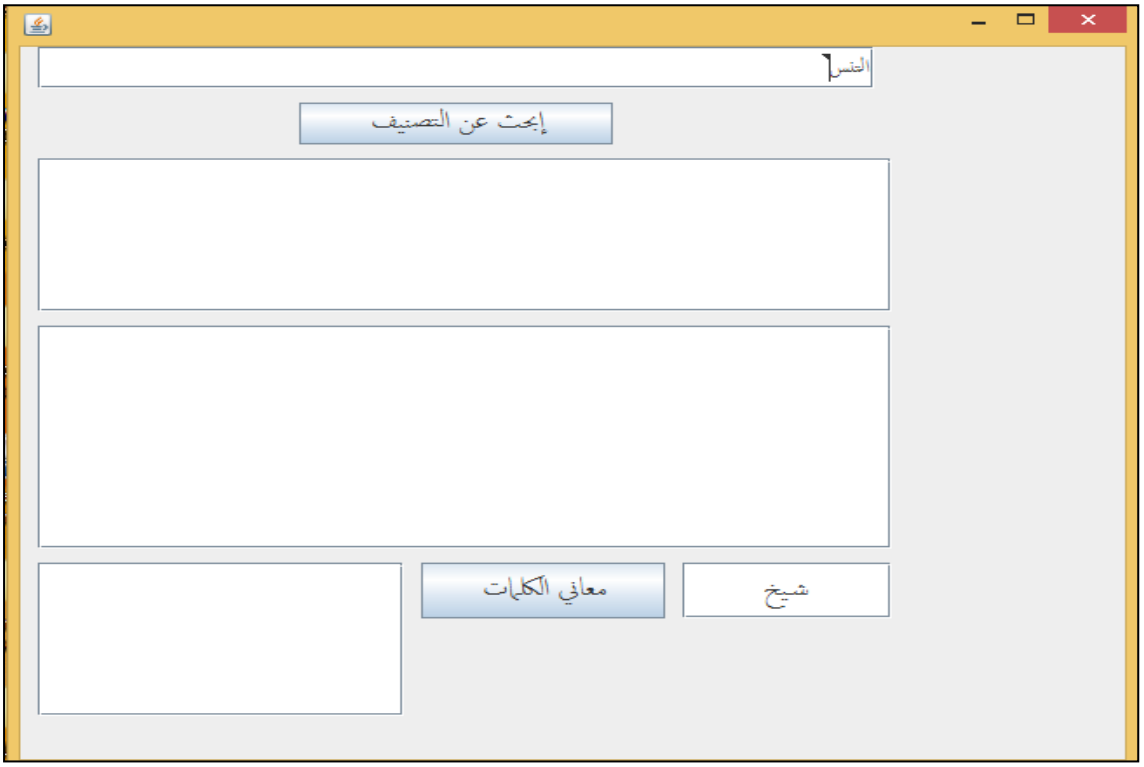

Figure 4.3 : Etape de Saisie

- après le clic sur la barre chercher la classe il nous affiche la catégorie du terme

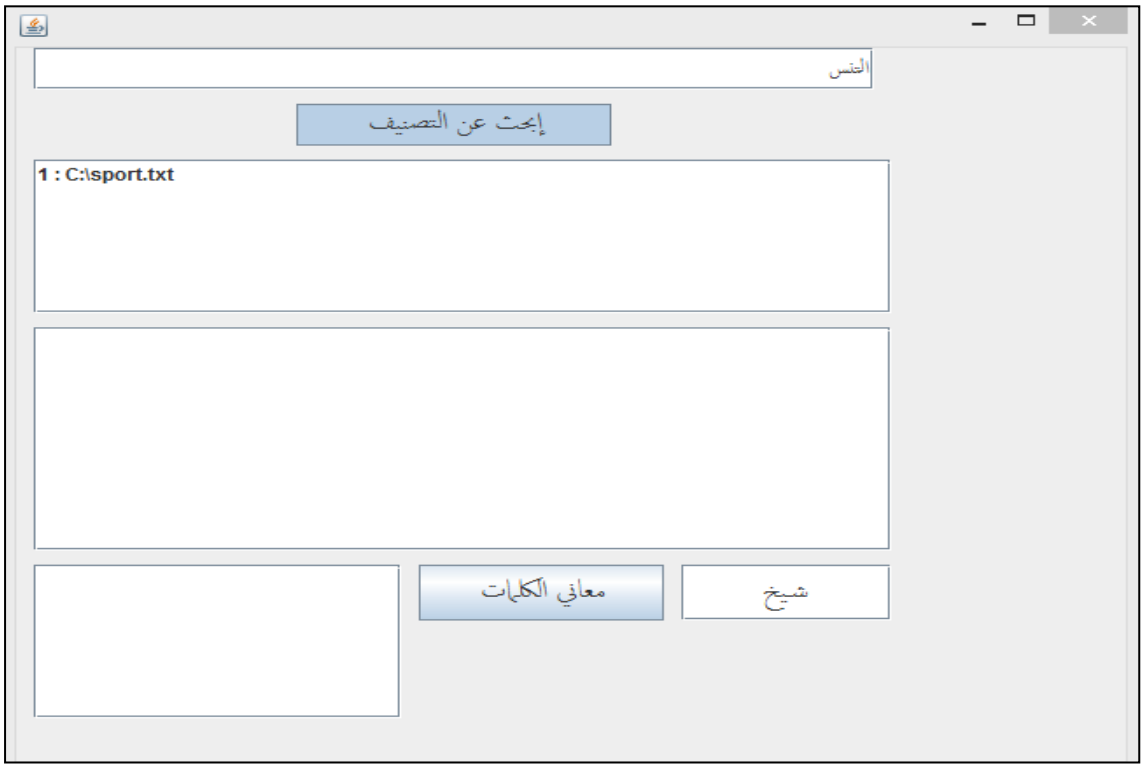

Figure 4.4 : Classification du terme

- Ensuite et si on le clic sur la catégorie classe le texte s'affiche

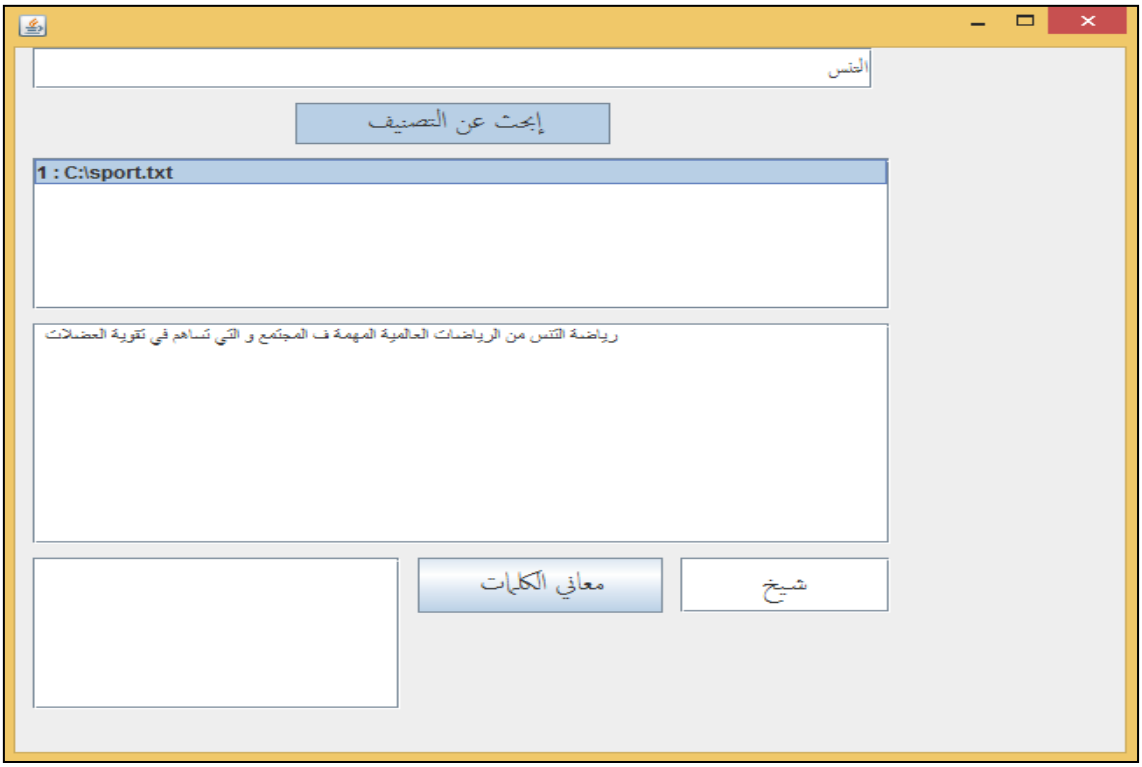

Figure 4.5 La classification du terme avec le texte

L'application permet l'extraction des synonyme a partir du dictionnaire Word net par l'explication des différentes terme et mots entres

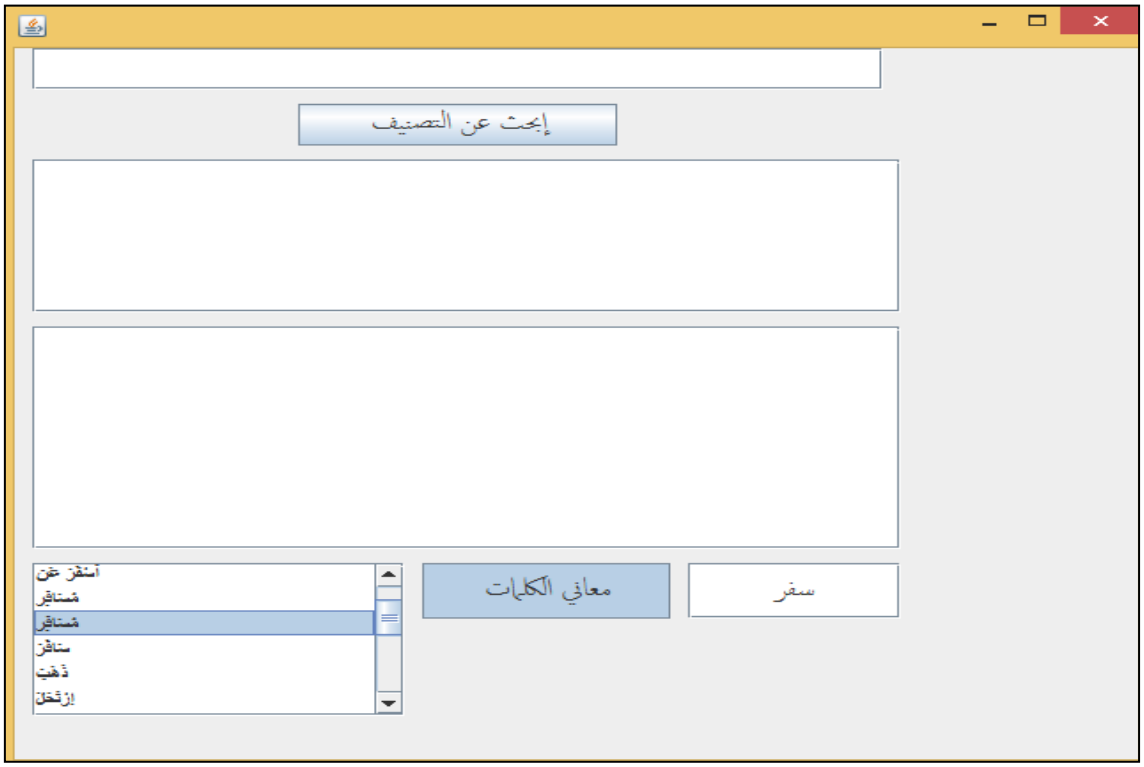

Figure 4.6 Synonyme des termes

4-Les différentes phases pour la réalisation

 Tokenization : cette étape est la production de séquence d'un segment séparé. La sortie est une liste des mots.

 Normalisation : Cette phase est représentée par Elimination des ponctuations, Elimination des voyelles, Elimination des non-lettres (chiffres, caractères spéciaux,…etc.).

 Eliminer les mots vides : cette étape constituée à la suppression les mots simples en utilisant une liste des mots appeler stop liste.

 Lemmatisation : Un des traitements les plus importants et une opération très difficile pour la langue arabe. en vue de la classification des documents est la racinisation ou lemmatisation des mots. Nous nous sommes basés sur l'ontologie Arabe WordNet pour prendre les lemmes de chaque mot.

5 L'algorithme utilisé :

J'ai choisi l'algorithme NB (naïf bayésien), pour construire notre modèle de prédiction qui nous permet d'associer des catégories aux différents textes.

$$P(Ci|D) = \frac{P(D|Ci) * P(Ci)}{P(D)}$$

Dans cette formule, P(Ci|D) représente la probabilité d'appartenance du document D à la catégorie Ci qui peut être également déterminée en évaluant la fréquence d'apparition des mots du document D qui sont associés à la catégorie Ci. P(D|Ci) est la probabilité selon laquelle, pour une catégorie donnée, les mots du document D sont associés à la catégorie Ci. P(Ci) est la probabilité qui associe le document D à la catégorie Ci indépendamment du contenu du document. P(D) est la probabilité propre du document D, toujours constant.

6 Discussion :

Dans cette expérience j'ai commencé avec une collection de documents où J'ai appliqué les étapes de traitement de texte à l'aide des ressources sémantiques (Word Net arabe dans mon cas) avec des mots de codage. j'ai également utiliser l'algorithme Naïf bayésienne avec WordNet arabe pour améliorer la qualité de la classification des textes arabes.

7- Conclusion :

Dans ce chapitre nous avons présenté les différentes étapes de classification qui jouent un grand rôle dans la conception qu'on a réalisé depuis la saisie du mot 'jusqu'à la phase de classification et en plus elle permis de trouver les différents synonyme des mots et termes on exploitant bien sure l'ontologie Arabe WordNet dans un système de classification des textes arabes..

Conclusion générale

L'utilisation de la langue arabe comme moyen de communication à travers le support informatique a été longtemps appréhendée avec beaucoup d'hésitation par la communauté scientifique, notamment celle du monde arabe où cet outil trouvera beaucoup d'utilisations importantes.En effet, la langue arabe et les différentes difficultés qui s'y rattachent, notamment le problème de l'ambiguïté issue de l'absence des voyelles,le problème de reconnaissance des formes fléchies (la langue arabe étant fortement flexionnelle) et le problème d'absence de travaux publiés sur l'extraction de l'information en langue arabe à travers l'utilisation de modèles statistiques du langage, s'ajoute à cela, la diversité des techniques et méthodes relatives au processus de classification qui pose un problème de choix, tout cela pose un énorme défi difficile à surmonter.

Malgré tout cela, nous avons osé nous aventurer dans ce domaine et on peut dire que, vu les résultats obtenus, nous pensons qu'on a quand même pu relever ce défi et par la même occasion apprendre beaucoup de nouvelles connaissances tout au long de la réalisation de ce travail telles que la classification automatique des documents et les techniques de recherche de l'information, le traitement automatique du langage naturel (en particulier, la langue arabe).

Toutefois, le sujet étant très vaste, il reste beaucoup à faire pour améliorer ce travail, on peut donc proposer comme perspectives, d'ajouter d'autres langues pour rendre le système multi-langues, d'étendre l'éventail des formes des mots arabes pris en compte par l'analyse morphologique, d'intégrer d'autres techniques et méthodes de classification supervisée, ainsi que toute autre idée jugée utile, réalisable et bénéfique.

Abstract

The texts classification is a task more or less complicated but frequently applied for: to assign a text with a category (class) according to its contents (categorization set of themes), to direct and filter other texts as being important or not important (e.g., filtering of spams), or even like integral part of the treatment of natural language NLP. The classical methods of classification are typically based on a model of representation known under the name “bag of words”, or each term and its derivatives are regarded as elements independent (to inform, information, data processing) by being unaware of any semantic relation between terms what leads to a failure for the algorithms of training during their application.

This project consists in using lexical ontology WordNet to capture any semantic relation between the terms what makes it possible to reduce the dimension of the space of representation of the texts on the one hand, and to improve the exactitude of categorization itself of another share.

Key words: document classification, supervised classification, algorithms, ontology, Word net.

ملخـــــــــــــص

ترتيب النصوص عملية أكثر ما يقال عنها أنها معقدة و لكنها غالبا ما تطبق من أجل إرفاق نص إلى فئة (صنف) معينة حسب المحتوى (تصنيف موضوعي) ، توجيه و فرز نصوص أخرى على أنها مهمة أو غير مهمة أو حتى أيضا على أنها جزء كامل من عملية معالجة اللغة.

المناهج الكلاسيكية للترتيب تؤسس نمطيا على نوع التمثيل و الذي يعرف تحت إسم (كيس الكلمات) أين تعتبر كل عبارة و مشتقاتها على أنهم عناصر مستقلة (استعلم ، استعلام ، إعلام آلي) متجاهلين في ذلك كل علاقة دلالية بين العبارات و الذي يوصل حتما إلى إخفاق خوارزميات التدريب خلال تطبيقها .

المشروع الآتي يرتكز على استعمال انطولوجيا وارد نات من أجل تحديد كل علاقة دلالية بين العبارات و الذي يسمح بتخفيض البعد الفراغي لتمثيل النصوص من جهة و إصلاح دقة الترتيب من جهة أخر

الكلمات المفتاح: ترتيب النصوص ، الترتيب المشرف، خوارزميات التدريب ، أنطولوجيا، وارد نات.

RESUME :

La classification des textes est une tâche plus ou moins compliquée mais fréquemment appliquée pour: assigner un texte à une catégorie (classe) selon son contenu (catégorisation thématique),orienter et filtrer d'autres textes comme étant importants ou non importants (e.g., filtrage de spams), ou même comme partie intégrale du processus de traitement du langage naturel NLP.

Les méthodes classiques de classification sont typiquement basées sur un modèle de représentation connu sous le nom "sac de mots", ou chaque terme et ses dérivés sont considérés comme des éléments indépendant (informer, information, informatique) en ignorant toute relation sémantique entre termes ce qui aboutit à un échec pour les algorithmes d'apprentissage lors de leur application.

Le présent projet consiste à utiliser l'ontologie lexicale WordNet pour capturer toute relation sémantique entre les termes ce qui permet de réduire la dimension de l'espace de représentation des textes d'une part, et d'améliorer l'exactitude de la catégorisation proprement dite d'une autre part.

Mots-clés : classification de documents, classification supervisée, algorithmes d'apprentissage, Ontologies, WordNet.

BIBLIOGRAPHIE

[1] CHAMI Djazia, Une plateforme orientée agent pour le data mining, En vue de l‟obtention du diplôme de Magister en informatique, Université HADJ LAKHDAR – BATNA, le : 2009.

[2] Maurice ROUX, ALGORITHMES DE CLASSIFICATION, Université Paul Cézanne Marseille, France, le : Juin 2006.

[3] MATALLAH Hocine, Classification Automatique de Textes Approche Orientée Agent, mémoire de Magister en informatique, Université AboubekrBelkaid-tlemcen, 2011.

[4]Radwan JALAM, Apprentissage automatique et catégorisation de textes multilingues, thèse de doctorat, université Lumière Lyon2, Année 2003.

[5] Yasmine Hanane Zeggane Mokhtar, Algorithmes d‟apprentissage pour la classification de documents, université de mostaganéme- algérie, licence 2009.

[6] https://fr.wikipedia.org/wiki/Ex6ploration_de_donn%C3%A9es

[7]HarragFouzi, Une approche de fouille des textes basée sur la classification et la segmentation thématique : Application au corpus des Traditions Prophétiques "Hadith", mémoire de doctorat d‟informatique, Université Ferhat ABBAS, Sétif, 2011.

[8]https://fr.wikipedia.org/wiki/Machine_%C3%A0_vecteurs_de_support

[9]https://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_artificiels

[10]https://fr.wikipedia.org/wiki/M%C3%A9thode_des_k_plus_proches_voisins

[11]https://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision

[12]https://fr.wikipedia.org/wiki/Classification_na%C3%AFve_bay%C3%A9sienne

[13] Laila Khreisat , A machine learning approach for Arabic text classification using N-gram frequency statistics, january 2009.

[14] Med El Amine Abderrahim, Vers une interface pour l‟enrichissement des requêtes en arabe dans un système de recherche d‟information.

[15] Karim Djelailia, AbdessalemKelaiaia, Hayat Farida Merouani Les machines à vecteurs supports dans la catégorisation de textes arabes.

[16] EL KHADIR LAMRANI, EL HABIB BEN LAHMAR, ABDELAZIZ MARZAK, ETUDES COMPARATIVE DES METHODES DE CLUSTERING DES TEXTES ARABES.

[17]TarekKanan and Edward A. Fox Automated Arabic Text Classification with P-Stemmer, Machine Learning, and a Tailored News Article Taxonomy.

[18] FarekLazhar, Identification d'opinions dans les textes arabes en utilisant les ontologies, thèse de doctorat, UNIVERSITÉ BADJI MOKHTAR–ANNABA, 2014.

[19] R. Neches, R.E. F ikes, T.Finin T, T.R.Ruber, R.Patil, T.Senator, W.R. Wartout, Enabling technology for knowledge sharing, AI magazine, 1991.

[21] N. Guarino, P. Giaretti, ontologies and knowledge bases: towards a terminological clarification. In towards very large knowledge bases, 1995.

[22] N. Aussenac-Gilles, B.Biébow, N.szulman, revisiting ontology design: a method based on corpus analysis, université Toulouse, 2000.

[**23]** Samia BOUARROUDJ, raisonnement sur une ontologie enrichie par des règles SWRL pour la recherche sémantique d‟images annotées, mémoire de magister, UNIVERSITE 20 AOUT 1955 Skikda, 2010.

[24] GASMI Mounira, Utilisation des ontologies pour l‟indexation automatique des sites Web en Arabe, Mémoire de MAGISTER, UNIVERSITE KASDI MERBAH OUARGLA, le : 27 mai 2009.

[25] Valéry Psyché, Olavo Mendes, Jacqueline Bourdeau, "Apport de l'ingénierie ontologique aux environnements de formations à distance", revue sticef.org, Volume 10,2003.

[26] Asuncion Gomez-Perez, "Ontological Engineering: a State of the Art", Expert Update. British Computer Society. Vol. 2. nº 3. pp. 33 – 43 (1999).

[27] Thomas Gruber, "A translation Approach to Portable Ontology Specifications", Knowledge Acquisition, 5(2), pp. 199 – 220.

[28] Guarino, N., «Formal Ontology and Information Systems», Formal Ontology in Information Systems. IOS Press, 1998.

[29] RiichiroMizuguchi, "Tutorial on ontological engineering - Part 1: Introduction to Ontological Engineering", New Generation Computing, OhmSha&Springer, Vol.21, No.4 (2003), pp.365 – 384.

[30] Moussaoui Kamal, FeredjDhiaElhak, Conception et développement d'un Outil de recherche sur le web à base d'agent, mémoire de master, UNIVERSITE KASDI MERBAH OUARGLA, 2013.

[31]Khadim DRAME, Contribution à la construction d‟ontologies et à la recherche d‟information : application au domaine médical, thèse de doctorat, université de bordeaux, le : 10 décembre 2014.

[32] https://fr.wikipedia.org/wiki/WordNet consulter le : 17/04/2017, 18 :30.

[33]Soraya Zaidi–Ayad, Une plateforme pour la construction d‟ontologie en arabe : Extraction des termes et des relations à partir de textes (Application sur le Saint Coran), Thèse Présentée en vue de l‟obtention du diplôme de DOCTORAT en Informatique, Université Badji Mokhtar Annaba, le : 29 juin 2013

[34] LahsenAbouenour, Karim Bouzoubaa, Paolo Rosso, Construction de l'ontologie Amine ArabicWordNet dans le cadre des systèmes Q/R.

[35] ABDERRAHIM Mohammed Alaeddine, Exploitation des Ontologies dans les Systèmes de recherche d'informations Arabes, thèse de doctorat, Université AboubakrBelkaïd Tlemcen, le : 25/ 02 / 2016.

[36] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller; Introduction to WordNet: An On-line Lexical Database (Revised August 1993)

[37] Christiane, Fellbaum, Wordnet – An Electronic Lexical Database, The MIT Press, Cambridge, Mass. ed. 1998.

[38] COLLINS A. M. & and Quillian M. R., 1969. « Retrieval Time From Semantic Memory ». Journal of Verbal Behavior and Verbal Learning 8 : 240-247.

[39] Mustapha BAZIZ, Indexation conceptuelle guidée par ontologie pour la recherche d'information, thèses de doctorat de l'université Paul Sabatier spécialité informatique, 2005

[40] William J. Black, Sabri Elkateb, Christiane Fellbaum, Musa Alkhalifa, Adam Pease, Horacio Rodríguez, Piek Vossen (2006) « Introducing the Arabic WordNet project Proceedings of the 3rd Global Wordnet Conference », Jeju Island, Korea, January, 2006. URL: http://www.lsi.upc.edu/~nlp/papers/fellbaum-alkhalifa-206.pdf.

[41] D. Farreres, « Creation of wide-coverage domain-independent ontologies». PhD thesis, Univertitat Politècnicade Catalunya, 2005.

[42] M. Diab, « The Feasibility of Bootstrapping an Arabic WordNet leveraging Parallel Corpora and an English WordNet». Proceedings of the Arabic language technologies and resources, nemlar, cairo 2004.

[43] N. Habash, and O. Rambow, « Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop ». In Proceedings of the 43rd Meeting of the Association for Computational Linguistics (ACL), 2005.