République Algérienne Démocratique et Populaire

Université SAAD DAHLAB - BLIDA1-

Faculté Des Sciences

Département De L'informatique

Mémoire de master en « informatique »

Option « traitement automatique de la langue »

Intitulé :

Développement d'un système d'extraction de concepts pour la langue arabe

Présenté par :

BELBAHRI Rim

BOUALBANI Hafsa

Devant le jury composé de :

Mr ABBACHE.A	Professeur à l'université de Chelef	Encadreur
Mm FARAH	Professeur à l'université de Blida1	Présidente
Mm Hadj Henni	Professeur à l'université de Blida1	Examinatrice
Mm Oukid.S	Professeur à l'université de Blida1	Promotrice

Année Universitaire 2018/2019

ملخص

بسبب ثروتها الصرفية والنحوية ، تعتبر اللغة العربية واحدة من أصعب اللغات للتعامل معها في مجال معالجة اللغة الطبيعية. ويرجع ذلك إلى الصعوبات المختلفة التي واجهتها في استخراج المعلومات ، والتي لم تشهد بعد نهجًا قياسيًا. هدفنا الأساسي هو تطوير نظام استخراج مفهوم عربي بسيط وفعال. تعتمد مقارنتنا على تحليل وورد نت العربي وتحليل المفهوم. حيث نختار مجموعات من الخصائص في السياق الرسمي من أجل بناء المفهوم الرسمي

الكلمات المفتاحية:

اللغة العربية ، المعالجة التلقائية للغة ، استخراج المعلومات ، ورد نت العربية ، تحليل المفهوم الرسمي

Résumé

Par sa richesse morphologique et syntaxique, la langue arabe est considérée parmi les langues les plus difficiles à traiter dans le domaine de traitement automatique de la langue. Cela est dû, notamment, aux diverses difficultés rencontrées dans l'extraction d'information (concept), qui n'a pas encore connu une approche standard. L'objectif de ce mémoire est de développer un système d'extraction de concept de la langue arabe simple et efficace basé sur wordnet arabe et l'analyse de concept formel.

Mots clé:

la langue arabe, traitement automatique de la langue, extraction d'information, wordnet arabe, analyse de concepts formels.

Abstract

Because of its morphological and syntactic richness, the Arabic language is considered as one of the most difficult languages to deal with in the field of natural language processing. This is due to various difficulties encountered in the extraction of information (consept) , which has not yet experienced a standard approach. Our primary goal is to develop a simple and effective Arabic concept extraction system. Our approche based on Arabic wordnet and consept formel analysis. Where we choose sets of properties in the formal context in order to construct the formal concept.

Keywords :

Arabic language , automatic language processing, information extraction, arabic wordnet, formal concept analysis.

Remerciement

Nous tenons d'abord à remercier ALLAH le tout puissant de nous avoir donné la volonté, l'amour du savoir et surtout le courage et la patience pour effectuer ce modeste travail.

Il nous tient à cœur d'exprimer toute notre reconnaissance a ceux qui au long de notre travail nous ont apporté leurs aides, leurs conseils, et leurs encouragements.

Nos sincères remerciements à Mr Abbache Ahmed professeur à l'université Chlef pour nous avoir encadrées et conseiller au cours de notre travail.

Nos sincères remerciements à notre chère professeur et promotrice Mm Oukid Saliha pour ces orientations et sa patience au cours de notre travail.

En fin, nous remercions nos amies pour leurs encouragements et leurs soutiens et tous ceux et celles qui de près ou de loin ont contribué à la réalisation de ce travail

Dédicace Rima

Avec l'aide de Dieu le tout puissant que j'ai pu arriver au terme de ce travail que Je tiens à dédier à :

A ma chère mère, la plus merveilleuse femme, je suis très reconnaissante à ses divers sacrifices,

Son soutien, et ses encouragements.

A mes sœur Lydia et Souad et Hayet et Nedjma et Nina et à mes frères : Mourad et Mohamed

A toute ma famille paternelle et maternelle

A ma copine avec qui j'ai préparé ce Modest travail Hafsa

A tous les étudiants da la promotion (2018/2019) de Informatique master TAL

A tous ceux ont attribué de près ou de loin à l'élaboration de ce Modes travail

Dédicace Hafsa

A la bougie de ma vie ma Mère

A mon père qui a toujours avec moi

Mes sœurs Karma et Hanaa et Soundoss

A mes frères Mohamed et Dirar et Akram et Diaa

A tout la famille Boualbani et la famille Elhabbes

A tout mes professeurs

A tout mes amis

Je dédié ce travail.

Sommaire

INTRODUCTION GENERALE 1

CHAPITRE 01 : Extraction d'Information et le traitement automatique de la langue arabe 3

1.1 Introduction :..... 4

1.2 L'extraction d'information..... 5

1.2.1 Les taches d'un système d'extraction d'information 6

1.2.2 Les approches d'EI 8

1.2.3 Les conférences MUC 9

1.2.4 Evaluation système d'extraction d'information 10

1.3 Traitement automatique de la langue arabe (TALA) 11

1.3.1 La langue arabe 11

1.3.2 Les niveaux de TALA 11

1.3.3 Particularité de la langue arabe 14

1.3.4 Les problèmes de traitement de la langue arabe 15

1.4 Conclusion 20

CHAPITRE 02 : Travaux connexes 21

2.1 Introduction :..... 22

2.2 Les traveaux..... 22

2.2.1 Travail 1 : Extraction de concepts guidée par le contexte [7] 22

2.2.2 Travail 2 :Extraction de concepts de texte basés sur WordNet arabe et analyse de concepts formels [9]: 24

2.2.3 Travail 3 :Une approche statistico-linguistique pour l'extraction de concept s sémantique [8]…… 26

2.2.4 Travail 4 : QuickUMLs : une approche rapide et non supervisée pour l'extraction de concepts médicaux [5]..... 27

2.2.5 Travail 5 : Combinaison des méthodes de reconnaissance d'entités nommées pour l'extraction de concept [2] 30

2.3 Etude comparative : 31

2.4 Conclusion : 33

Chapitre 03 : Conception et Réalisation..... 34

3.1 Introduction :..... 35

3.2 Définition de concept : 16

3.3 Outils et formalisme utilisés : .....**Error! Bookmark not defined.**

3.3.1 WordNet Arabe (ArabicWordNet AWN) :..... 16

3.3.2 Analyse de concepts formels(ACF) : 18
3.4 Architecture global du système : 35
3.5 L'approche proposée : 36
3.6 Architecture de système: 37
3.7 Conclusion : 40
Chapitre 04 : Implémentation et Evaluation 41
4.1 Introduction: 42
4.2 Implémentation : 42
4.2.1 Outils et Environnement de développement : 42
4.2.2 Description : 43
4.2.3 Déroulement : 44
4.2.4 Exemple de traitement d'un texte : 47
4.3 Evaluation : 48
4.4 Conclusion : 51
CONCLUSION ET PERSPECTIVES 52
Bibliographie 53

Liste des Figures

Figure 2-1:Les modules de système[7] 23
Figure 2-2:Les étapes d'extraction de concepts [9] 25
Figure 2-3:L'architecture détaillée de système [8] 26
Figure 2-4:Aperçu des identificateurs d'entités nommées [2]. 30
Figure 3-1:Les relations dans wordnet [4] 17
Figure 3-2:Les synsets de mot'مأوى'. 18
Figure 3-3:L'architecture globale de système d'extraction de concepts. 35
Figure 3-4:Architecture détaillée de premier système. 37
Figure 3-5:Architecture détaillé de deuxième système. 38
Figure 3-6:Exemple de prétraitement d'une phrase 40
Figure 4-1:Interface principale de système. 44
Figure 4-2: Normalisation d'un texte 45
Figure 4-3: La segmentation d'un texte 46
Figure 4-4: Le résultat selon ACF 46
Figure 4-5: Le résultat selon AWN 47
Figure 4-6: Graphe de rappel pour ACF et AWN 50

Liste des tableaux

Tableau 1-1:Extraction d'information sur un article du journal [3] 6
Tableau 1-2:Conférence MUC [3]..... 10
Tableau 1-3:La racine des mots..... 12
Tableau 1-4:Les schèmes des mots 12
Tableau 1-5:Les lemmes des mots 12
Tableau 1-6:Quelques phrases acceptées par la grammaire de la langue arabe 13
Tableau 1-7:Représente déférents écritures de la lettre <<qaf>>..... 14
Tableau 2-1:Etude comparative entre les travaux. 32
Tableau 3-1:Représentation de contexte formel[15] 19
Tableau 4-1: Les résultats rappel et précision des deux approche du système..... 49

INTRODUCTION GENERAL :

Ces dernières années, en raison de la quantité croissante de supports et de données électroniques arabe disponibles, qui contient des informations précises, la nécessité d'analyser automatiquement les documents électroniques s'est accrue.

Pour indexer des informations ou chercher une information pertinente, il y a des problèmes de complexité de l'indexation automatique en général et manque d'outils d'indexation en langue arabe. Cela nécessite l'extraction automatique de concepts en langue arabe.

Objectifs

L'objectif du travail consiste à concevoir, réaliser et à évaluer un système d'extraction de concept qui est capable d'extraire automatiquement à partir d'un texte écrit en Arabe, les différentes concepts. Le système reçoit en entrée des documents textuels et produit en sortie une liste des concepts.

Problématique

Ces dernières années sont marquées par une augmentation énorme de la quantité d'information électronique arabe et dont l'accès à des informations pertinentes est devenu de plus en plus complexe et le besoin de développer des applications d'aide à la lecture est devenu incontournable.

La réalisation d'un système d'extraction de concept constitue un domaine à part entière se trouvant à la croisée de traitement automatique de la langue et de la recherche d'information.

Les travaux dans ce domaine existent déjà pour d'autres langues comme l'anglais ou le français, malheureusement pour l'arabe les travaux sont rares et les choses ne font que commencer. Nous essayons donc de contribuer dans ce sens en proposant une méthode d'extraction de concepts pour les textes arabes.

Plan de mémoire

Notre travail est organisé en quatre chapitres : le premier concerne l'extraction d'information et le traitement automatique de la langue arabe et les difficultés rencontrées. Dans le deuxième chapitre nous nous intéressons à des travaux connexes sur l'extraction d'information et nous effectuons une étude comparative entre ces travaux afin de choisir nos approches. Ensuite un troisième chapitre est dédié à la description des deux approches utilisées ainsi que l'architecture de chaque approche et en termine avec leur implémentation, les tests et l'évaluation dans le quatrième chapitre.

1

CHAPITRE 01 : Extraction D'Information et le traitement automatique de la langue arabe

1.1 Introduction :

Le Traitements Automatique des Langues (TAL) est une discipline qui associe étroitement linguistes et informaticiens. Elle repose sur la **linguistique**, les **formalismes** (représentation de l'information et des connaissances dans des formats interprétables par des machines) et l'**informatique**. Le TAL a pour objectif de développer des logiciels ou des programmes informatiques capables de traiter de façon automatique des données linguistiques.

Pour traiter automatiquement ces données, il faut d'abord expliciter les règles de la langue puis les représenter dans des formalismes opératoires et calculables et enfin les implémenter à l'aide de programmes informatiques.

Les principaux domaines du TAL sont :
- le traitement de la parole .
- la traduction automatique.
- la compréhension automatique des textes.
- la génération automatique de textes.
- la gestion électronique de l'information et des documents existants (GEIDE).

Depuis les débuts du Traitement Automatique du Langage (TAL) dans les années 60, la compréhension automatique de textes est l'objet de nombreuses recherches vise à saisir le sens global d'un document. L'extraction d'information (EI) est l'un des problèmes les plus intéressants dans le traitement du langage naturel (TLN). L'EI permet de stocker dans une base de données factuelle l'information jugée pertinente en vue de traitements ultérieurs

L'EI consiste à extraire des informations précises de documents et à les structurer sous une forme prédéfinie. Il s'agit en général de remplir des formulaires donnant certaines caractéristiques concernant des entités ou des événements évoqués dans les textes ainsi que des relations entre ces entités et ces événements. Le formulaire est constitué d'une liste d'attributs auxquels le système doit faire correspondre une liste de valeurs pour chaque texte analysé.
Dans le vaste éventail des traitements automatiques de documents textuels, il sera commode de situer l'EI comme un niveau intermédiaire entre la recherche documentaire et la compréhension automatique (au sens de l'intelligence artificielle (IA)).

1.2 L’extraction d’information :

L’extraction d’information est le nom donné à tous processus qui structure et combine les données. Il désigne une technologie récente qui vise à extraire et à structurer automatiquement un ensemble d’informations précises apparaissant explicitement ou implicitement dans un ou plusieurs documents textuels écrits en langue naturelle. Cette nécessité s’est accrue ces vingt dernières années avec l’essor considérable de la masse de documents disponibles au format électronique qu’il faut gérer afin d’extraire ou de filtrer les informations pertinentes parmi toutes celles contenues dans ces documents.
La sortie finale du processus d’extraction d’information varie dans tous les cas, cependant il peut être transformé afin de remplir un certain type de base de données.

Exemple : un article de journal

Texte

لقي 11 شخصا بورقلة بينهم 03 نساء وأستاذة جامعية وطفلة لم يتجاوز عمرها سنتين، حتفهم في حادث مرور مروع، اقتصرت له الأبدان، بينما أصيب 28 راكبا بينهم عسكري بجروح متفاوتة الخطورة لا يزالون يتلقون العلاج بمستشفى محمد بوضياف وسط المدينة، حالة 06 منهم خطيرة، واثنان في قاعة الإنعاش.

الحادث المأسوي وقع الأحد في الساعة السادسة صباحا بالطريق الوطني رقم 56 الرابط بين عاصمة الولاية ومدينة تقرت بالنقطة الكيلومترية رقم 60، نجم عن احتكاك حافلة لنقل المسافرين من نوع "هيجر" تابعة لمؤسسة خاصة، كانت قادمة من العاصمة يقودها شيخ، بشاحنة ذات مقطورة يقودها شاب، مما استدعى تدخل مصالح الحماية المدنية معززة بـ08 سيارات إسعاف و04 شاحنات إطفاء وتسخير 40 عونا.

Informations	
Attribut	Valeur
Nombre de victimes	11
Nombre de blessés	28
Lieu	ورقلة
Date	الأحد في السادسة صباحا
Cause	عن احتكاك حافلة لنقل المسافرين من نوع هيقر تابعة لمؤسسة خاصة ,كانت قادمة من العاصمة يقودها شيخ ,بشاحنة ذات مقطورة يقودها شاب

Tableau 1-0-1: Extraction d'information sur un article du journal [1]

1.2.1 Les taches d'un système d'extraction d'information :

Définition :

L'extraction d'informations a pour tâche d'identifier les occurrences d'une classe d'entités, de relations et d'événements spécifiés au préalable dans des textes en langage naturel, ainsi que l'extraction des propriétés (arguments) pertinentes des entités, relations ou événements identifiés. Les informations à extraire sont pré-spécifiées dans des structures définies par l'utilisateur, appelées modèles (ou objets), chacune consistant en un certain nombre d'emplacements (ou d'attributs), qui doivent être instanciés par un système EI lors du traitement du texte. Les remplacements de fentes sont généralement les suivants: chaînes du texte, l'une des valeurs prédéfinies ou une référence à un modèle d'objet généré précédemment. Une façon de penser à un système EI est en termes de population de base de données, puisqu'un système EI crée une représentation structurée (par exemple, des enregistrements de base de données) d'informations sélectionnées tirées du texte analysé.

Les composants trouvés dans les systèmes d'EI d'aujourd'hui reflètent largement les tâches définies dans ces conférences. Les tâches de la dernière conférence [1][2], MUC-7, en 1998 (les plus difficiles dans la série) ont été les suivantes :

- Reconnaissance des entités nommées.
- Détection de la coréférence.
- Reconnaissance des éléments du formulaire.

- Reconnaissance des relations.
- Reconnaissance des scénarios (« scenario template »).

Reconnaissance des entités nommées

Cette tâche consiste à repérer toutes les formes linguistiques bien identifiées, à l'instar des noms propres de personnes, d'organisations, de lieux, etc. mais aussi les expressions temporelles (dates, durées…), les quantités (monétaires, unités de mesures, pourcentages…) et à leur affecter une étiquette sémantique choisie dans une liste prédéfinie [2].

Détection de la coréférence

Cette tâche consiste à repérer les groupes nominaux et les pronoms personnels co- référents et à les baliser dans les textes. Par exemple, dans « En 1963, Warda El-Djazairia épouse Djamel Kesri, un des fondateurs de l'ancienne sécurité militaire. Après son exil en Égypte, elle se rendait de moins en moins à son domicile à Alger. », La résolution des coréférences devrait relier « Elle » à « Warda El-Djazairia » [2].

Reconnaissance des éléments du formulaire

Cette tâche, qui repose sur les deux tâches précédentes, consiste à associer des informations (descriptions, informations complémentaires) aux entités reconnues. Elle associe en fait de l'information descriptive, généralement sous la forme de groupes nominaux, aux entités précédemment identifiées. Cette information descriptive correspond à un attribut de l'entité concernée [2].

Reconnaissance des relations

La reconnaissance des relations s'attache à identifier un certain nombre de relations, le plus souvent binaires, entre les entités extraites précédemment. Ainsi, dans l'exemple précédent, cette tâche permet de repérer une relation de mariage entre les entités personnes « Warda El-Djazairia » et « Djamel Kesri » [2].

1.1.1 Reconnaissance des Scenarios

Cette tâche relie entre les entités et les relations précédemment reconnues des descriptions d'évènement relatif au domaine étudié. Les différents traits complémentaires, telles que la localisation spatiale et temporelle sont également associés. La reconnaissance des scénarios est une tâche particulièrement difficile. Elle dépend des résultats des étapes précédentes et possède donc un score plus faible, dépendant de la composition de leurs résultats [2].

1.2.2 Les approches d'EI *:*

Dans la littérature, différents travaux sont utilisés pour extraire les termes du texte de script de deux manières: analyse statistique, analyse linguistique. L'analyse statistique est basée sur l'étude des contextes d'utilisation et la répartition des termes dans les documents. L'analyse linguistique exploite les connaissances linguistiques, telles que les structures morphologiques ou grammaticales des termes. Les autres œuvres combinent ces deux approches et forment une approche dite "hybride ou mixte"

1.2.2.1 Approche symbolique :

L'Approche statistique (symbolique) se base principalement sur les cooccurrences et la fréquence de mots dans le texte pour d'extraire des termes. L'idée principale est que si deux mots coexistent souvent dans les mêmes contextes, alors ils peuvent être regroupés. Cette idée a été réalisée avec succès dans plusieurs travaux. L'avantage de cette approche est qu'elle ne nécessite pas de connaissances externes (comme Word net par exemple). Par conséquent, elles restent indépendantes et applicables dans tout domaine. De plus, les associations entre les termes (synonymie, polysémie) sont partiellement prises en compte. Dans de nombreuses expérimentations, elle a obtenu des performances meilleures que les techniques classiques de RI mais reste limitée pour le traitement de grandes collections de documents. Les méthodes statistiques sélectionnent les termes en fonction de leur distribution dans le corpus [1].

1.2.2.2 Approche linguistique :

Cette approche s'appuie sur des règles linguistiques et utilise des ressources linguistiques (phonétiques, morphologiques, syntaxiques et sémantiques) [1].

Les auteurs commencent par extraire des phrases candidates qui peuvent être pertinentes pour les documents dans lesquels elles apparaissent. Pour cela, ils appliquent la lemmatisation et la partie de la parole marquage afin qu'ils puissent identifier la catégorie grammaticale des mots. Ensuite, un modèle linguistique est appliqué, de sorte qu'une expression doit commencer et se terminer par un substantif ou un adjectif et peut contenir d'autres noms, adjectifs, prépositions ou articles entre les deux. Le résultat final est une liste de phrases et leur fréquence d'occurrence. L'étape suivante est la sélection des phrases, où différentes stratégies sont discutées. Une des stratégies est de sélectionner les phrases avec la fréquence la plus élevée de l'occurrence en couvrant le nombre maximum de documents récupérés. Une autre stratégie consiste à utiliser l'analyse de fréquence, bien que la restriction de l'ensemble des phrases candidates à ceux qui contiennent un ou plusieurs des termes de la requête d'origine. La dernière stratégie attribue des valeurs plus élevées aux expressions qui se produisent plus fréquemment dans l'ensemble de documents récupérés que dans l'ensemble de la collection, un peu semblable à TF-IDF. Le reste du document traite du processus de clustering ayant les caractéristiques mentionnées comme base [2].

1.2.2.3 Approche hybride non supervisée :

Une approche linguistique et statistique pour l'extraction de mots-documents scientifiques. Dans cette approche, Zervanou[1] commence par un prétraitement linguistique des textes, à savoir son étiquetage en partie de la parole et l'identification de zones spécifiques des documents, tels que le titre, l'abrégé, l'introduction, les conclusions, les remerciements et Références. Ensuite, l'étape suivante consiste à identifier les phrases-clés candidates, en modes de règles morphosyntaxiques prédéfinis. Afin de réduire la variation des résultats après l'application d'une mesure statistique, l'auteur propose la normalisation du texte. Afin de réduire la variation morphologique, il utilise le lexique WordNet pour obtenir les lemmes de chaque phrase-clé candidate, tandis que pour les variations orthographiques, telles que les phrases composées par rapport aux mots composés non-coupés, elles sont traitées par des techniques de correspondance des règles.

Enfin, l'auteur applique la mesure de valeur 'C' pour obtenir un score pour un multi-mot. Cette métrique de la valeur 'C'est essentiellement la multiplication de la fréquence d'occurrence d'une expression par sa longueur. Comme pour d'autres approches linguistiques, l'utilisation d'outils spécifiques à la langue et, dans ce cas, de règles linguistiques pour obtenir des lemmes et identifier différents concepts orthographiquement écrits similaires, impose une dépendance linguistique. Par exemple, WordNet n'est pas disponible pour de nombreuses langues et n'est pas complet, même pour l'anglais (dans le sens d'inclure toutes les combinaisons possibles ou les relations). Cela peut impliquer un rappel plus bas que souhaité. En outre, l'utilisation de la longueur d'un terme dans le calcul de la valeur C peut impliquer la suppression de mots clés plus courts, tels que la RAM ou ROM dans un article sur la mémoire de l'ordinateur.

1.2.3 Les conférences MUC :

Les conférences de compréhension du message (Message Understanding Conference MUC) constituent un forum permettant d'évaluer et discuter des progrès réalisés dans le domaine du traitement du langage naturel. Chaque conférence est procédée d'une évaluation formelle des systèmes d'analyse de texte mis au point pour exécuter une tache partagée, conçue par le gouvernement en consultation avec les participants à l'évaluation du monde de la recherche [3].

Les recherches actuelles en EI ont été influencées par les conférences MUC qui sont déroulées entre 1987 et 1998 faisaient partie du programme TIPSTER[2] financés par DARPA[3]. Ce programme comportait trois tâches : la détection des documents, l'extraction d'information, et le résumé de texte.

[1] Maître de conférences au sein du groupe Systèmes d'information de l'université de technologie d'Eindhoven
[2] Le programme TIPSTER : https://www-nlpir.nist.gov/related_projects/tipster/overv.htm
[3] DARPA :Defense Advanced Research Projects Agency.

(MUC 1,1987) et (MUC 2,1989)	Traitement et analyse d'un petit nombre de message naval
(MUC 3 ,1991) et (MUC4 ,1992)	Le but était d'analyser un grand nombre d'article de presse complet dans des domaines vaste. Ils ont requis le traitement des textes journalistiques en anglais sur le terrorisme en Amérique latine.
(MUC 5,1995)	Création des joint-ventures et circuits électronique
(MUC 6,1996)	A traité les changements de personnels de direction de grandes entreprises.
(MUC 7,1998)	L'analyse de textes journalistiques rapportant des craches d'avion et des tirs de missiles

Tableau 0-2:Conférence MUC [1]

1.2.4 Evaluation de système d'extraction d'information :

La mise au point de méthodes d'évaluation des systèmes de Traitement Automatique du Langage (TAL) apparaît aujourd'hui comme une composante importante, quoique difficile, de la recherche dans ce domaine .Même si cette évaluation présente bien sûr des limites, la spécification des formats de sortie et des critères d'évaluation ainsi que la réalisation d'outils d'évaluation automatique constituent par elles-mêmes des avancées certaines pour la recherche [1].

L'évaluation de système d'extraction d'information peut se faire avec les mesures suivantes:

Le rappel (R) : est une évaluation de la couverture du système. Il mesure la quantité de réponses pertinentes d'un système par rapport au nombre de réponses idéales [1].

$$R = \frac{\text{Nombre d'entités correctes détectés}}{\text{Nombre d'entités manuellement identifiées}}$$

La précision (P) : est une évaluation du bruit du système. Elle mesure la proportion des réponses correctes parmi l'ensemble des réponses fournies par le système [1].

$$P = \frac{\text{Nombre d'entités correctes détectés}}{\text{Nombre d'entités détectés}}$$

Le F-mesure (F) : C'est la moyenne harmonique de la précision et du rappel qui mesure la capacité du système. À donner toutes les solutions pertinentes et à refuser les autres, une mesure populaire qui combine la précision et le rappel est leur pondération.

$$\mathrm{F} = \frac{2(\mathrm{P}.\mathrm{R})}{(\mathrm{P}+\mathrm{R})}$$

E-mesure(E) : F-mesure paramétrique.

Permet d'attribuer un ordre de préférence entre le rappel et la précision.

$$E = \frac{(1+\beta^2)PR}{\beta^2 P+R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R}+\frac{1}{P}}$$

- β=1 : même poids précision et rappel
- β>1 : privilégie la précision au rappel
- β<1 : plus d'importance au rappel.

1.3 Traitement automatique de la langue arabe (TALA) :

Le traitement automatique de la langue arabe est une discipline en pleine expansion, dans laquelle on voit de plus en plus de recherches et de technologies se soucier des spécificités de cette langue et proposer des outils nécessaires au développement de son traitement automatique.

1.3.1 La langue arabe :

La langue arabe est considérée comme5ème langue courante utilisée dans le monde elle fait partie de la grande famille des langues sémitiques. La langue arabe est une langue difficile à maitriser dans le domaine de traitement automatique de la langue par ses propriétés morphologique et syntaxique.

1.3.2 Les niveaux de TALA :

- La morphologie :

Il étudie la construction et la transformation des unités lexicales selon le sens voulu [4]. La racine, le schème et le lemme sont des principaux concepts qui décrivent la majorité des noms et des verbes tel que :

- **La racine** : c'est la base de toutes les formes des verbes et des noms. Les racines sont à l'origine de la plupart des mots arabes. Il s'agit d'une séquence de consonnes qui peuvent être trouvées dans tous les mots qui en sont dérivés. Elle est un élément important dans les langues dérivationnelles.
  Exemple :

Mot	تستذكرون	حروب	أعرابي
Racine	ذكر	حرب	عرب

Tableau 1-3:la racine des mots

- **Le schème** : un schème est une forme prédéfinie qui caractérise une classe de verbe ou de nom (des schèmes verbaux et des schèmes nominaux). Les schèmes sont des modèles avec des structures différentes qui sont appliqué à la racine pour créer un mot. À la même racine, nous pouvons appliquer différents schèmes pour avoir des mots différents avec des significations différentes. La langue arabe comprend environ 150 schèmes. Dans le tableau suivant nous donnons les mots générés (verbes et noms) à partir de la racine (كتب/ktb) selon le schème appliquée :

mot	كتب	كاتب	كتابة	مكتبة	مكتوب
Schème	فعل	فاعل	فعالة	مفعلة	مفعول

Tableau 1-4:Les schèmes des mots

- **Le lemme :** c'est l'entrée lexicale dans un lexique ou dans un dictionnaire, il peut être analysé comme une racine insérée dans un schème. Il est l'intersection entre une forme graphique et un sens. La connaissance du lemme ou de couple (racine, schème) permet de déduire les différentes formes fléchies d'un verbe ou d'un nom. Ainsi, la dérivation morphologique est décrite sur une base morphosémantique d'une même racine décrivent différentes unités lexicales selon des schèmes.

Exemple :

Mots	الحروب	ناقة	استغنم
Lemme	حرب	ناقة	غنم

Tableau 1-5:Les lemmes des mots

- La syntaxe :

  Elle étudie la structure correcte des phrases en analysant l'ordre des unités lexicales. La définition de la phrase, qui convient à l'anglais et du français, ne peut s'appliquer à tous les schémas de phrase de la langue arabe. Cette dernière dispose, en plus de la phrase verbale, de deux schémas de phrases dans lesquels l'élément central n'est pas un verbe et dans lesquels aussi, il peut n'y avoir aucune marque explicite de temporalité. Il s'agit en arabe de la phrase nominale et de phrase locative (composée d'un group prépositionnel ou d'un circonstanciel) [4].

  Le tableau suivant illustre quelques exemples de phrases acceptées par la grammaire de la langue arabe alors que leur équivalence en langue française donne des phrases incomplètes voire erronées :

Phrase en langue arabe	Phrase acceptable par la grammaire arabe	Phrase traduite en français	Phrase acceptable par la grammaire française	Phrase corrigé
الكتاب فوق الطاولة	✓	Le livre sur la table	X	Le livre est sur la table
الطالب ذاهب إلى الجامعة	✓	L'étudiant allant à l'université	X	L'étudiant est en train d'aller à l'université
في المدرسة معلمون	✓	A l'école instituteurs	X	Il y'a des instituteurs à l'école

Tableau 1-6:quelques phrases acceptées par la grammaire de la langue arabe

- La Sémantique :

L'analyse sémantique tente de découvrir de façon plus générale le sens des mots, des phrases ou des textes entiers. C'est la phase la plus laborieuse pour les machines, et pour cette raison elle reste encore assez peu employée.

L'absence de voyelles peut générer des défauts de sens dans le traitement automatique, par exemple, le mot (العلم) isolé peut avoir plusieurs interprétations (la science ou drapeau) alors que voyellé seraال) pour la science et (المُع)pour le drapeau). Les outils qui opèrent cette analyse font souvent appel à de gigantesques thésaurus (comme

ArabicWordnet pour l'arabe), permettant de classer chaque terme dans une arborescence de concepts pour déterminer les thèmes dominants d'un texte, ainsi qu'à des algorithmes complexes permettant d'évaluer les relations entre les différentes idées d'un texte donné. [5]

1.3.3 Particularité de la langue arabe :

- La représentation morphologique de l'arabe est assez complexe en raison de la variation morphologique et du phénomène d'agglutinement[4], les lettres changent de forme selon leur position dans le mot (isolé, initiale, médiane et final)

Exemple :

Isolé	Initial	médiane	Final
ق	قرآن	القرآن	غسق

Tableau 0-7:représente déférents écritures de la lettre <<qaf>>

- La langue arabe s'écrit de droite à gauche, le majuscule n'existe pas, et les signes diacritiques représentent les voyelles courtes [5], ces voyelles ne sont pas des lettres de l'alphabet, ce sont des signes diacritique qui se rajoutent aux consonnes (lettres) ses signes sont :

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

[4]Une agglutination désigne le rassemblement d'éléments mis en contact

- 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 [5][14].

1.3.4 Les problèmes de traitement de la langue arabe :

La langue arabe est une langue complexe surtout pour le traitement automatique, parmi les problèmes de traitement de cette langue en peut citer :

➢ La segmentation :

Pour le traitement automatique de la langue arabe la segmentation est une étape fondamentale, son rôle est de découper le texte en unités d'un certain type qu'on aura défini et repéré probablement. En effet, l'opération de segmentation d'un texte consiste à délimiter les segments de ses éléments de base qui sont les caractères, en éléments constituants différents niveaux structurels tel que : paragraphe, phrase, syntagme, mot graphique mot-forme, morphème etc [5].

La langue arabe n'est pas appuyée principalement sur les signes de ponctuation, il est à noter que ses derniers ne sont pas utilisés de façon régulière dans les textes arabes actuels, et même dans le cas où ils y figurent, ils ne sont pas gérés par des règles précises d'utilisation. Par ailleurs, nous pouvons trouver tout un paragraphe ne contenant aucun signe de ponctuation à part un point à la fin de ce dernier. Ainsi, il convient de noter que la présence des signes de ponctuation ne peut guider la segmentation comme c'est le cas du français ou de l'anglais.

D'autre particularités de cette langue s'ajoutent et posent un problème lors de la segmentation du texte en phrases nous en citons :

- L'absence de majuscules qui marquent le début d'une nouvelle phrase.
- L'ambigüité des signes de ponctuation et de certaines particules comme "و" et,"ف" qui peuvent dans certains cas, ne plus jouer le rôle de déclencheurs d'une nouvelle phrase.

➢ L'agglutination :

La plupart des mots arabes sont composés par agglutination d'éléments lexicaux de base (proclitique + base + enclitique). Par exemple, la détermination peut s'exprimer par agglutination de l'article ال?/al/ avant le mot (المالية?/almaleya/ (« financement »)ou par agglutination d'un pronom personnel après celui-ci (ماله/malohu/ (« son argent »).Dans

toute perspective de traitement automatique, le problème est donc de décomposer le mot en ces différentes parties. Cette décomposition nécessite des connaissances de niveau supérieur en cas d'ambiguïtés (si le mot accepte plusieurs segmentations).

➢ L'étiquetage grammatical:

L'étiquetage grammatical (tagging en anglais) est l'opération qui consiste à attribuer à chacun des mots d'un texte sa catégorie (nom, verbe, adjectif, article défini, etc.).

La difficulté de l'étiquetage grammatical s'amplifie lorsque les textes visés se présentent sous leur forme non pas voyellé, mais partiellement voyellée ou encore totalement non voyellée, ce qui correspond au cas le plus courant [5].

1.4 Définition de concept :

Les connaissances modélisées (dans l'ontologie) portent sur des objets auxquels on se réfère à travers des concepts. Le concept représente un ensemble d'objets et leurs propriétés communes. Les concepts correspondent aux abstractions pertinentes d'un segment de la réalité (le domaine du problème, retenues en fonction des objectifs fixés et de l'application envisagée (pour l'ontologie) [6].

Un concept est composé de trois parties :

- Une notion : elle correspond à la sémantique du concept, elle est définie à travers ses propriétés et ses attributs. Elle est appelée intention du concept.
- Un ensemble d'objets : il correspond aux objets définis par le concept, il est appelé extension du concept. Les objets sont les instances du concept.
- Un (ou plusieurs) terme(s) : les termes permettent de désigner le concept. Ces termes sont aussi appelés labels de concept.

1.5 WordNet Arabe (ArabicWordNet AWN) :

WordNet arabe est une ressource lexicale pour l'arabe moderne standard basé sur le WordNet largement utilisé de Princeton pour l'anglais (Fellbaum, 1998). WordNet arabe (AWN) est basé sur la conception et le contenu de Princeton WordNet (PWN) universellement accepté et sera directement mappable sur PWN 2.0 et EuroWordNet (EWN), permettant ainsi la traduction du niveau lexical en anglais et des dizaines d'autres langues. [7]

Les AWN ont une structure de thésaurus, il est organisé autour de la structure des synsets dont chaque synset est un concept [6] qui est un ensemble de synonymes et de lien décrivant la relation avec d'autres synsets. Ces relations peuvent être définies comme suit :

- la synonymie (similaire).

- antonyme (opposé).
- hypéronymie (super concept) / hyponymie (sous-concept) (aussi appelé Is-A hiérarchie / taxonomie).
- méronymie (partie de).
- holonymie (has-a).

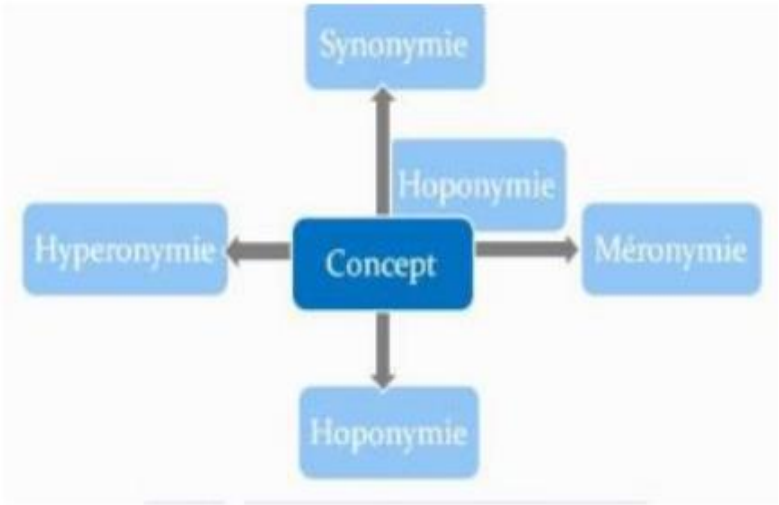

Figure 3-1:Les relations dans wordnet [6].

Les relations sémantiques entre les synsets diffèrent selon la catégorie grammaticale. Chaque mot peut appartenir à un ou plusieurs synsets, et à une ou plusieurs catégories du discours. L'AWN est donc un réseau lexical où les nœuds sont les synsets et les relations entre les synsets sont les arcs.

Enfin, AWN est une ressource pour la langue arabe générale disponible en ligne. Il a actuellement 11269 synsets et 23 481 mots.

Exemple : Les synsets de mot'مأوى'

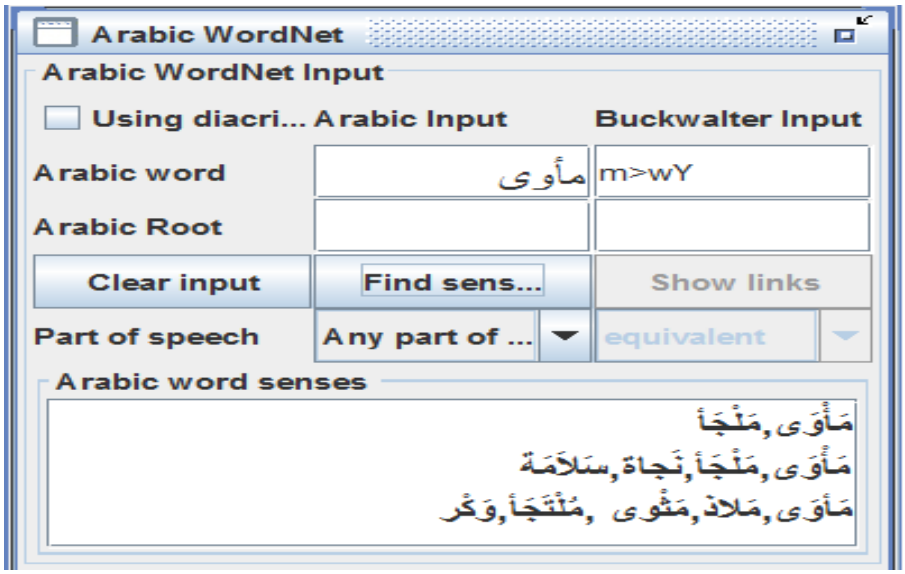

Figure 0-2:les synsets de mot'مأوى'.

1.6 Analyse de concepts formels(ACF) :

L'Analyse de Concepts Formels (ACF) (appelée aussi Analyse formelle de Concepts (AFC)) est un formalisme mathématique pour l'analyse de données, la représentation et la visualisation de connaissances. L'idée de base de l'ACF est d'extraire des concepts regroupant des objets et leurs propriétés/attributs à partir de données et de construire une hiérarchie à partir de ces concepts. [8]

1.6.1 Contexte formel :

Un contexte formel est un triplet (G,M,R) où G est un ensemble d'objets, M est un ensemble de propriétés/attributs et R est une relation binaire entre G et M. Un couple (g,m) ∈ R (également noté gRm) signifie que l'objet g ∈ G possède la propriétés m ∈ M.

Un contexte formel peut être représenté sous la forme d'un tableau à deux dimensions où les lignes correspondent aux objets et les colonnes aux propriétés. Les cases du tableau sont remplies suivant la présence/absence de la propriété, autrement dit si l'ième objet g est en relation R avec le jème alors la case à l'intersection de la ligne i et de la colonne j contient "×", sinon la case est vide. La table suivant donne un exemple de contexte formel :[8]

G= {D1, D2…..D10}

M= {C ,E,COMP :YES,COMP :NO}

	C	E	COMP:YES	COMP:NO
D_1	×		×	
D_2			×	
D_3	×		×	
D_4				×
D_5		×		×
D_6	×		×	
D_7			×	
D_8		×		×
D_9				×
D_{10}	×	×		×

Tableau 0-8:représentation de contexte formel[15] .

1.6.2 Concept formel :

À partir d'un contexte formel, nous calculons le concept formel.

Nous avons un couple E, I, où E est l'ensemble maximal d'objets (appelé extension) possédant toutes les propriétés de I et I est l'ensemble maximal des propriétés (appelé intension) partagé par tous les objets de E.

Dans un treillis de concepts les données sont structurées sous forme de concepts. Un concept peut être vu comme une classe d'objets (l'extension du concept) caractérisée par un ensemble de propriétés (l'intension du concept)[8].

Les liens généralisation/spécialisation

Dans un treillis de concepts les concepts sont ordonnés selon deux critères duaux liés à leurs extensions et à leurs intensions. Les concepts les plus généraux sont situés en haut du treillis alors que les concepts les plus spécifiques sont situés en bas du treillis. Les liens entre les concepts peuvent être interprétés comme des généralisations ou des spécialisations entre les classes représentées par les concepts. En effet, un parcours ascendant des concepts d'un treillis se traduit à chaque étape par la diminution progressive du nombre d'attributs dans les intensions des concepts et l'augmentation progressive du nombre d'objets dans leurs extensions [8].

1.7 Conclusion

Des nouvelles perspectives s'ouvrent dans la recherche en EI .Il faut souligner tout l'intérêt de ce domaine de recherche. En premier lieu, les retombées industrielles dans le domaine de l'informatique documentaire, sous les formes variées, sont probablement proches. Mais son impact pour la recherche fondamentale ne doit pas être sous-estimé.

De nombreuses applications dépendent de l'extraction automatique de la structure à partir de données non structurées pour obtenir de meilleurs moyens d'interrogation, d'organisation et d'analyse des données reliant le monde structuré et non structuré. Partant de recherches menées dans la communauté de langage naturel sur les systèmes de base de reconnaissance d'entités nommées, le sujet fait désormais appel à une véritable communauté de chercheurs couvrant l'apprentissage automatique, les bases de données, le Web et la recherche d'informations. Il existe actuellement beaucoup de travail sur divers aspects du problème d'extraction d'informations, notamment des modèles statistiques et basés sur des règles, des cadres et des architectures pour la gestion des pipelines d'extraction, l'optimisation des performances, la gestion des incertitudes, etc.

2

CHAPITRE 02 :

Travaux connexes

2.1 Introduction :

Dans ce chapitre nous présentons les articles qui vont nous aide de choisir une méthode pour notre projet. Ce chapitre se décompose en 5 travaux:

- extraction de concepts guidée par le contexte [9] .
- extraction de concepts de texte basés sur WordNet arabe et analyse de concept formel [10].
- une approche statistico-linguistique pour l'extraction de concepts sémantique [11].
- combinaison de la reconnaissance d'entités nommées pour l'extraction des concepts [12].
- QuickUMLs : une approche rapide et non supervisée pour l'extraction de concepts médicaux [13].

Les limites de chaque méthode de ces articles et les hypothèses fixant notre objectif permettent ensuite de mieux comprendre vers quel choix nous avons orienté notre méthodologie. A la fin nous allons choisi une méthode pour l'implémenter.

2.2 LES TRAVEAUX

2.2.1 Travail 1 : Extraction de concepts guidée par le contexte [9]

2.2.1.1 *Introduction :*

Cet article présente une méthode d'extraction de concepts ontologiques utilisant un algorithme de clustering non supervisé et guidé par le contexte à partir de pages Web , cette méthode est basée sur une approche unifiée intégrant des dimensions complémentaires pour l'acquisition de connaissances conceptuelles. Ils ont exploité les caractéristiques structurelles des documents HTML.La réalisation du web sémantique dépend de la construction des ontologies. Tel que les ontologies permettent de représenter un ensemble de concepts formellement définis, acceptés par une communauté d'utilisateurs, selon les domaines et les besoins applicatifs. Ils ont expérimenté sur un corpus du domaine portant sur le tourisme. Les premiers résultats obtenus montrent que la prise en compte du contexte des termes guidant le clustering améliore considérablement la pertinence des concepts extraits.

2.2.1.2 *Architecture du système de travail:*

Architecture de ce système est composée principalement de trois étapes : une étape de prétraitement, une étape de traitement et une étape de formalisation et d'évaluation, ainsi le système a comme tâche la recherche, la découverte et la structuration des connaissances conceptuelles à partir des pages Web en vue de construire une ontologie de domaine. Les composants de système assistent l'ingénieur de connaissances lors de la constitution et du traitement du corpus, la représentation des données, l'analyse du corpus, la découverte et la structuration des connaissances en produisant une ontologie de domaine ou en la raffinant.

Les acteurs du système sont l'utilisateur et l'expert, tel que l'utilisateur pourra intervenir lors des deux étapes le prétraitement et le traitement, l'expert de domaine pourra évaluer les classes de mot, l'ontologie. La figure suivante présente les modules du système :

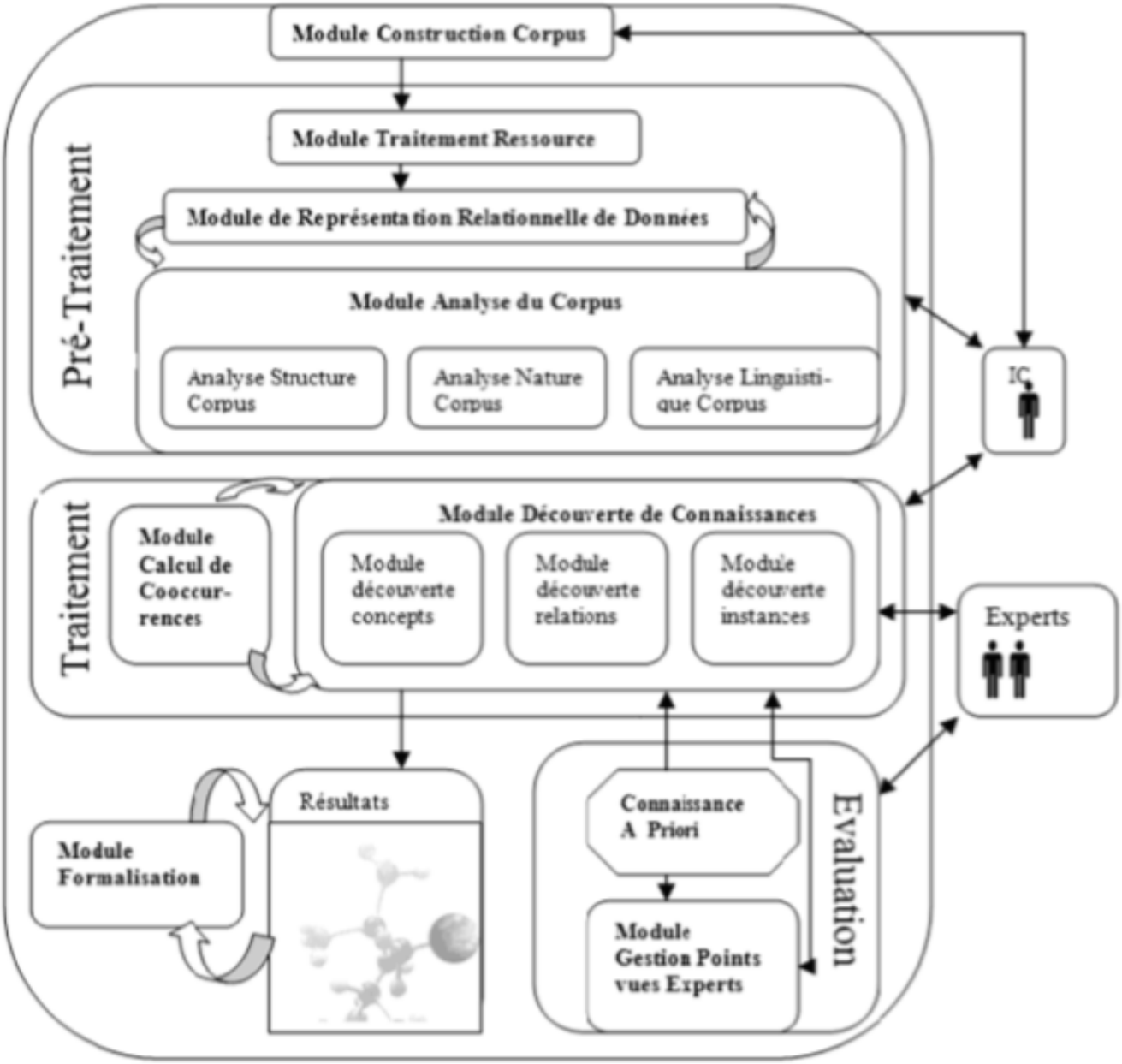

Figure 2-1:Les modules de système[9]

2.2.1.3 *Résultat :*

Afin d'évaluer le modèle contextuelle, ils ont appliqué deux définitions de contextes sur le corpus ; le premier contexte est un contexte statique (cherche les co-occurrents d'un mot dans un espace de 10 mots), le deuxième contexte se base sur leur hiérarchie contextuelle (appliqué uniquement sur les mots appartenant aux balises clefs définies) après l'application du clustering sur les deux définition , ils expérimentent différentes alternatives de nombre de classes allant de 20 à 400

Pour évaluer et présenter les résultats, ils ont défini six critères : la distribution des termes, la pondération de paires de termes, la similarité de paires de termes, les concepts extraits, l'interprétation sémantique et le degré de généralité des concepts extraits. Concernant la distribution des termes, avec le contexte statique, ils obtiennent 74 classes parmi lesquelles il existe une classe contenant 55% des termes. Alors que pour le contexte basé sur la structure HTML, seuls 13% des termes initiaux sont regroupés dans la même classe.

2.2.2 Travail 2 : Extraction de concepts de texte basés sur WordNet arabe et analyse de concepts formels [10]:

2.2.2.1 *Introduction :*

Pour améliorer la capacité des systèmes de recherche d'information il est nécessaire de concevoir et de développer des méthodes basées sur un traitement de texte sémantique, permettant de choisir les termes appropriés, qui peuvent représenter sémantiquement le contenu de ce texte .Cet article présente une méthode permettant l'extraction de concepts qui représentent le contenu sémantique d'un texte arabe. Ces concepts sont extraits de l'arabe WordNet (AWN), qu'ils vont appliquer ensuite pour leur analyse de concept formel, afin de produire un ensemble de concepts plus réduits et plus pertinents. Dans cet article les chercheurs ont adopté une nouvelle approche consistant à intégrer l'aspect sémantique au cours du processus d'indexation. Il s'agit de ce que nous appelons l'indexation sémantique ou l'indexation conceptuelle. Cet article présente une méthode permettant d'extraire des concepts qui vont représenter le contenu sémantique d'un texte arabe. Ces concepts sont extraits de la ressource lexicale AWN, selon les termes présents dans le texte, puis d'un processus de la désambiguïsation est appliquée aux termes ambigus pour identifier exactement leur sens approprié et finalement ils appliquent pour ces concepts l'analyse de concept formelle qui vise à augmenter la capacité de l'approche à produire un ensemble de concepts plus réduite et plus pertinente, et en particulier étant donné une meilleure représentation du contenu textuel sémantique.

2.2.2.2 *Approche proposée:*

Cet article présente une approche qui vise à extraire des concepts qui à représentent un texte arabe, ces concepts pouvant être utilisés ultérieurement en tant que descripteurs sémantiques pour la phase d'indexation. Cette approche est basée d'une part sur AWN et d'autre part sur le formalisme d'analyse de concepts formels (FCA). Le système est divisé en trois étapes:

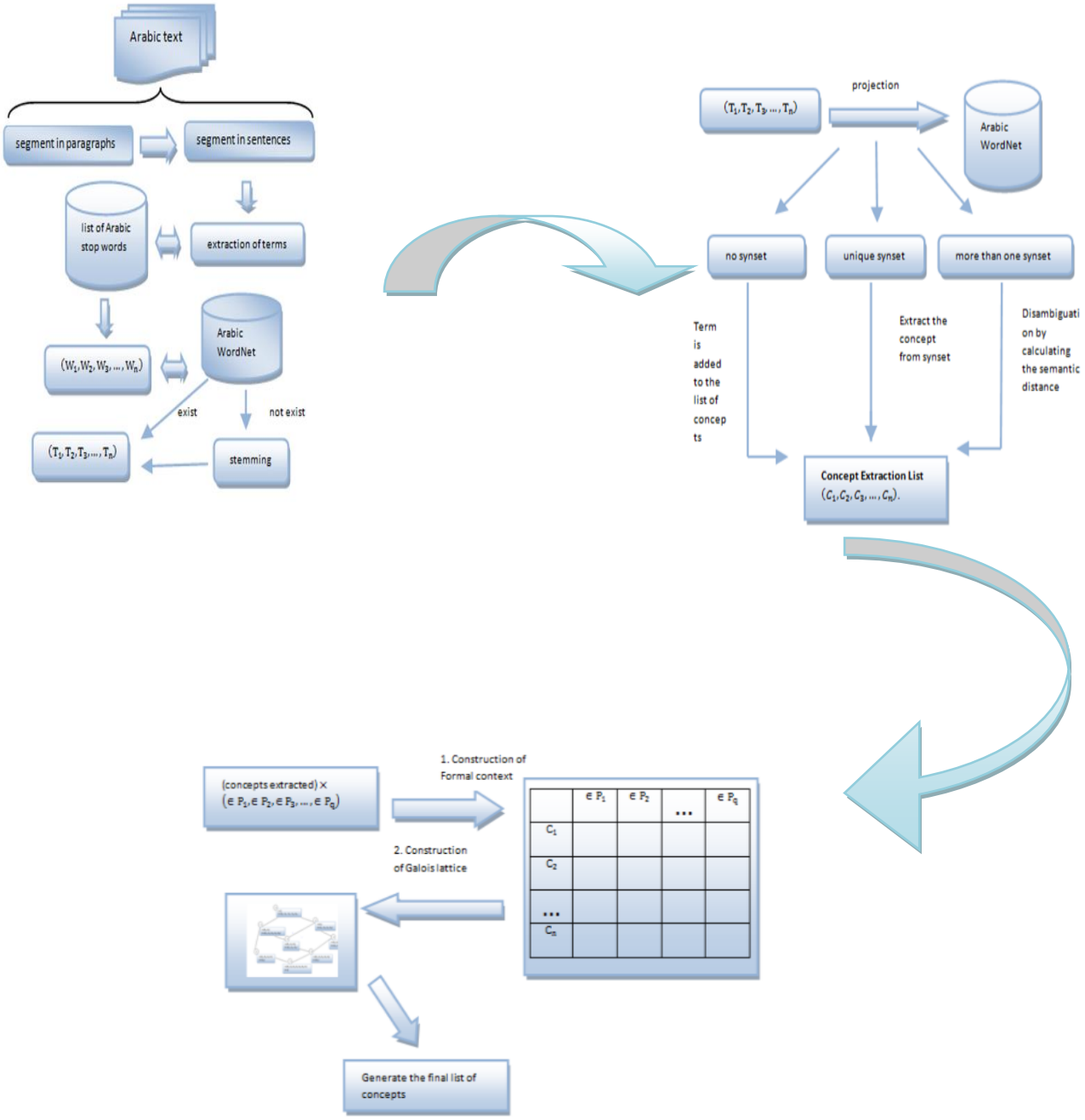

Figure 2-2:Les étapes d'extraction de concepts [10]

2.2.2.3 Analyse de concepts formels :

Analyse de concepts formels (ACF) est un processus qui permet de découvrir tous les regroupements possibles d'objets ayant des propriétés communes. Et aussi une méthode mathématique basée sur la théorie de l'ordre, utilisé pour l'analyse des données

2.2.3 Travail 3 : Une approche statistico-linguistique pour l'extraction de concepts sémantique [11] :

2.2.3.1 *Introduction :*

Cet article présente un système de compréhension automatique de l'énoncé. L'architecture est basée sur une approche statistico-linguistique qui suppose que la compréhension de l'énoncé est une traduction de langage naturel vers un langage conceptuel .Pour cela ils proposent une nouvelle méthode d'extraction de concepts sémantique basée sur la recherche documentaire LSA.

2.2.3.2 *La méthode :*

D'après Didierjean[5], si on considère un énoncé de problème S constitué de N objet. Comprendre cet énoncé se résumer a en trois étapes, une analyse descendante de l'énoncé S, afin de trouver les différent objets. Puis une identification des relations établies entre ces objets. Enfin, une analyse ascendante depuis ces relations qui aboutira à la compréhension de l'énoncé.

L'architecture du système de compréhension de l'énoncé est organisée en deux étapes : Etude et Test .Elle est montrée ci-dessus :

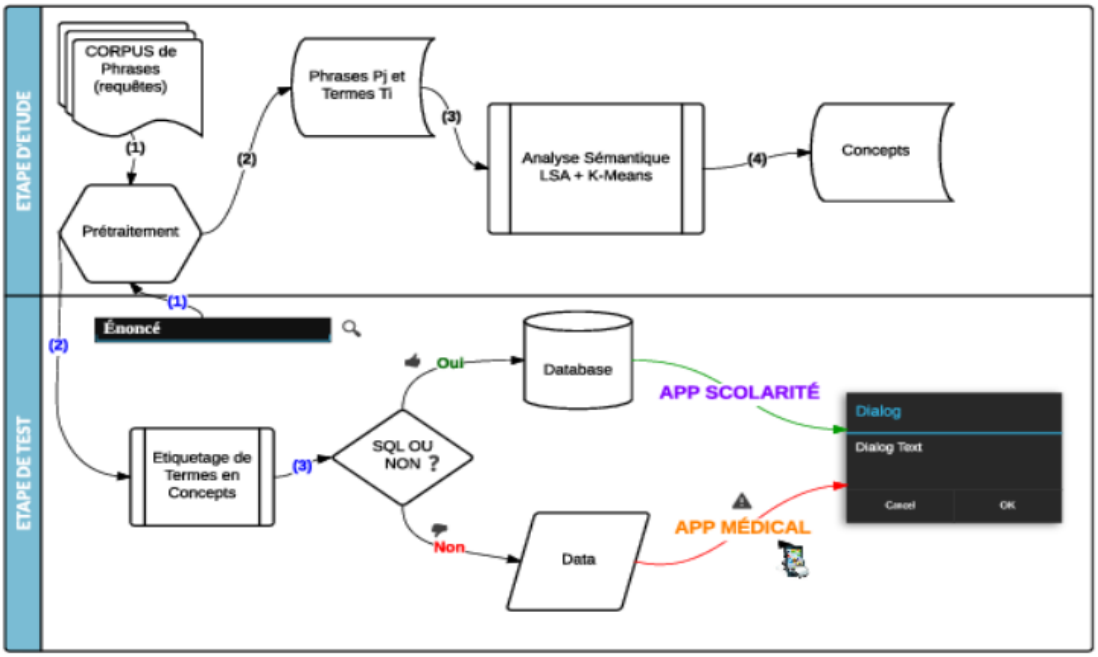

Figure 2-3:L'architecture détaillée du système [11]

[5] Est un patronyme français typique du département des Vosges en région Lorraine.

Etape d'étude :

1. Analyse structurelle.
2. Analyse lexicale.
3. Extraction de concept.
4. Représentation LSA

Etape de test :

1. Langage générique.
2. Gestionnaire de termes hors vocabulaire.

2.2.3.3 *Evaluation :*

Pour apprécier la qualité du système, ils ont considérés en plus deux applications : application de consultation scolaire et application de diagnostic médical.

La validation des résultats trouvées est assurée par la mesure F1 défini par :

$$F1 = \frac{2 * precesion * rappel}{précesion + rappel}$$

Le taux d'extraction des concepts avec l'application de consultation est de l'ordre de 51.85% sur un corpus de 21 requêtes. Alors que le taux de compréhension sur 60 requêtes est de l'ordre de 73.33% qui est beaucoup mieux que celui basé sur l'approche de recherche par mots clés qui est de l'ordre de 68.33%. D'autre part, ils ont atteint un taux de reconnaissance (extraction de concepts) de l'ordre de 61.11%, mais concernant l'étape de compréhension la tâche était irréalisable à cause de la richesse linguistique du domaine médical et en plus de la grammaire générative des requêtes (symptôme) du patient qui ne peut pas être crée.

2.2.4 Travail 4 : QuickUMLs : une approche rapide et non supervisée pour l'extraction de concepts médicaux [13].

2.2.4.1 Introduction :

Dans ce travail, le système présenté QuickUMLS s'appuie sur une correspondance approximative avec les termes en UMLS pour extraire les concepts médicaux à partir de texte non structuré. La méthode proposée est nettement plus rapide que les autres systèmes tels que MetaMap et cTAKES.

2.2.4.2 Méthodologie :

Soit un dictionnaire S un ensemble de chaînes, C collection de concepts, C: C → S une carte associant un concept de la collection à une ou plusieurs chaînes du dictionnaire. Soit un document d représenté comme une suite de tokens {d1,. . . , dn}

CPMerge :

CPMerge détermine $Y_{x,a}$comme suit: premièrement, il considère les trigrammes de caractères comme des entités, par exemple : la chaîne $x = \text{“}tumor\text{”}$ est associée à l'ensemble des fonctionnalités suivantes: X = {## t, #tu, tum, umo, mor, or #, r ##}, où le signe dièse indique le début ou la fin d'une chaîne.

Ensuite, il calcule la taille minimale et maximale de l'ensemble de caractéristiques Y, de toute chaîne $y \in S$ pouvant avoir au moins τ caractéristiques communes avec x. Lorsque la similarité de Jaccard est utilisée,$min\,|\,\mathrm{Y}\,| = [\alpha \cdot |\,\mathrm{X}\,|\ ]$ et $max\,|\,\mathrm{Y}\,| = [\,|\,\mathrm{X}\,|\,/\,\alpha]$. Le nombre minimal de caractéristiques qui se chevauchent τ dépend également de la fonction de similarité utilisée; dans ce cas, $\tau = [(\alpha\,(|\,X\,| + |\,Y\,|)\,/\,(1 + \alpha)]$. Par exemple, pour $x = \text{“}tumor\text{”}$, $|\,X\,| = 7$, ainsi, toute chaîne $y \in S$ telle que $Jaccard\,(x, y) >= 0.7$ doit avoir entre $[0.7 \cdot 7] = 5$ et $[7\,/\,0.7] = 10$.

Une fois les longueurs minimale et maximale déterminées, CPMerge obtient, pour chaque $l \in \{min\,|\,Y\,|, min\,|\,Y\,| + 1 \ldots, max\,|\,Y\,|\}$, l'ensemble $Y_{x,a,l}$de chaînes de longueur l ayant plus de τ caractéristiques communes à x. Cela est fait en joignant les listes de publication de chaque entité dans X et en conservant les chaînes qui apparaissent dans plus de τ listes. Pour des raisons d'efficacité, chaque liste d'écriture est partitionnée en ensembles contenant toutes les chaînes de même longueur. Ainsi, CPMerge peut rejoindre le sous-ensemble de chaque liste de publication contenant des chaînes de longueur l sans accéder à la liste complète. L'ensemble de toutes les correspondances approximatives avec le dictionnaire de la chaîne x est obtenu en considérant l'union de $Y_{x,a,l}$pour toutes les valeurs acceptables de l.

QuickUMLS :

Étant donné un document d de longueur n, un seuil de similarité α, et une taille de fenêtre w, QuickUMLS génère efficacement, pour chaque token $di \in d$, toutes les séquences possibles de tokens $d_{ij} = \{d_i, \ldots, d_j\}, j \in \{i, \ldots, i + w - 1\}$ .Ensuite, un ensemble d'heuristiques est utilisé pour déterminer si dij est une séquence valide de token; Si tel est le cas, CPMerge est utilisé pour identifier les chaînes similaires à dij dans S. Une fois que le sous-ensemble de toutes les chaînes de correspondance possibles $Z_{d,a} = U_{d_{ij}}(Y_{d_{ij},a})$ est déterminé, QuickUMLS sélectionne le sous-ensemble le plus approprié $Z'_{d,a}$de chaînes de sorte qu'il n'y ait pas de chevauchement entre l'ensemble des concepts extraits. En détail, QuickUMLS procède comme suit. Tout d'abord, il marque et obtient une partie des étiquettes de parole pour chaque token. SpaCy5 a été utilisé pour accomplir les deux tâches. Ensuite, ∀i, j ∈ {1,. . . , n}, j ≤ (i + w - 1), il génère toutes les séquences valides possibles de tokens. Une séquence de tokens d_{ij} est valide si :

• dij contient au plus w tokens. Un token pourrait être un mot, une ponctuation ou un nombre.

• dijne couvre pas les phrases.

• Le premier ou le dernier token de la séquence (d_i etd_j, respectivement) ne sont pas une conjonction, une adposition ou un déterminant tel que déterminé par les balises de partie de parole extraites. Ces tokens sont conservés tant qu'ils apparaissent dans une séquence. Cela a été déterminé comme étant une bonne stratégie pour manipuler des cordes telles que «difficulté à marcher».

• Le premier token de la séquence (d_i) n'est pas une ponctuation.

• dij n'est pas un mot d'arrêt ni un nombre si dij ne contient qu'un seultoken (c'est-à-dire, i = j).

Ensuite, CPMerge est utilisé pour rechercher dans S des chaînes dont la similarité avec d_{ij} est supérieure ou égale à α et les ajouter à$Z_{d,a}$. Étant donné l'ensemble $Z_{d,a}$ de toutes les séquences possibles dans d pouvant être associées aux concepts en C, QuickUMLS sélectionne un sous-ensemble de chaînes sans recouvrement qui optimise la somme des scores de similarité avec les séquences de tokens dans d. En d'autres termes, pour deux séquences superposées d_{ij}etd_{pq}, d_{ij} est ajouté à $Z'_{d,a}$ si pour tout s ∈S , $strsim\ (d_{ij}, s) > strsim\ (d_{pq}, s)$. En cas d'égalité (c'est-à-dire qu'il existe s ∈ S tel que $strsim\ (d_{ij}, s) = strsim\ (d_{pq}, s)$), la séquence la plus longue est choisie.

2.2.4.3 Evaluation :

cTAKES a la meilleure précision

• Le meilleur rappel de QuickUMLS est proche de cTAKES lorsque sim = 1.0

• F1: QuickUMLS = 0.63, cTAKES = 0.61, MetaMap = 0.48

Les résultats sont similaires à i2b2

• cTAKES a toujours la meilleure précision, le meilleur rappel QuickUMLS

• F1: QuickUMLS = 0,72, cTAKES = 0,68 *, MetaMap = 0,61 *

• QuickUMLS est 2-26 fois plus rapide que cTAKES

2.2.5 Travail 5 : Combinaison des méthodes de reconnaissance d'entités nommées pour l'extraction de concept [12] :

2.2.5.1 Introduction :

La reconnaissance d'entité nommée (NER) est une tâche clé de l'extraction de la sémantique à partir de texte. Dans cet article la méthode proposée pour les NER en micro-postes, conçue pour combiner les annotations générées par les outils NER existants afin de produire des résultats plus précis que les seuls outils de saisie. Ils combinent les identifiants entité nommées en utilisant des techniques d'apprentissage automatique, à savoir l'arbre de décision et la forêt aléatoire en utilisant l'algorithme C4.5.

Les dernières années ont vu une croissance significative de l'interaction avec les médias sociaux. Les personnes peuvent interagir via Internet de presque n'importe où et à tout moment. Ils peuvent partager leurs expériences, leurs pensées et leurs connaissances instantanément, et dans des dimensions de masse. Le moyen le plus simple et probablement le plus populaire d'interaction sur le Web consiste à utiliser des micro-postes, de courts messages texte postés sur le Web.

2.2.5.2 L'approche proposée :

L'idée de combiner les identifiants entité nommée consistait à utiliser des techniques d'apprentissage automatique pour créer un modèle de classification, qui serait formé aux fonctionnalités décrivant le texte du micro-message ainsi qu'aux annotations produites par les identifiants entité nommée impliqués. Ils ont utilisé le jeu de données d'apprentissage pour construire le modèle et le jeu de données test pour l'évaluer et le comparer à d'autres outils de reconnaissance des éléments de réseau. Sept identifiants NE sur huit ont été choisi de combiner, en utilisant différentes méthodes.

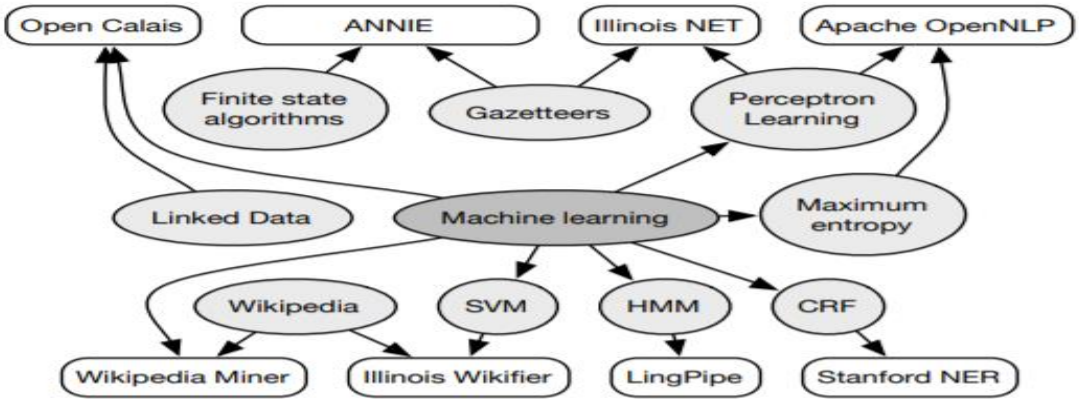

Figure 2-4:Aperçu des identificateurs d'entités nommées [12].

Combinaison de dispositifs de reconnaissance des EN :

- Reconnaissance de NE de base
- Transformer les annotations NE en vecteurs
- Prétraitement des données de formation
- Formation et évaluation des modèles

2.2.5.3 Evaluation :

L'évaluation sur un ensemble de données standard montre que l'approche proposée surpasse les méthodes de NER sous-jacentes, ainsi que l'outil de reconnaissance de pointe NE spécialement formé sur les micro-données. Au meilleur des connaissances, à ce jour, l'approche proposée obtient le meilleur score F1 sur le jeu de données # MSM2013.

Les techniques d'apprentissage automatique les plus performantes étaient les arbres aléatoires de forêt et de décision basés sur l'algorithme C4.5

2.3 Etude comparative :

Dans le tableau suivant, nous comparerons les différents travaux cités dans le deuxième chapitre par rapport au corpus, l'approche utilisée et la méthode appliquée et les résultats de chaque article.

Les trois travaux [12] [13] [11] ont donnés des bons résultats avec un taux de réponse environ 60%, Cependant, Quick a obtenu le meilleur résultat

				Critère de test		
Référence	**Approche**	**Corpus**	**Méthode**	**Rappel**	**Précision**	**Résultat**
Loubna et Maria **[7]**	Approche unifiée intégrant des dimensions complémentaires pour l'acquisition de connaissances conceptuelles	Domaine de tourisme	Extraction de concepts ontologiques			
Nadia et Faouzia et Habib(/2015) **[9]**	Approche sémantique	Arabe wordnet(AWN)	Extraction de concepts représentant le contenu sémantique d'un texte arabe.			
Mohamed et Amar et Rachida (2015) **[8]**	Approche statistico-linguistique		Extraction de concepts sémantique basée sur la recherche documentaire LSA.			
Luca et Nazli (2016) **[5]**	Approche non supervisé	-défi i2b2/VA	Extraction de concepts médicaux à partir de texte non structuré	67%	75%	63%
		-thyme		87%	77%	72%
		-examens des Médicaments		47%	60%	48%
Stefan (2016) **[2]**	Approche statistique	MSM	combiner les annotations générées par les outils NER	61.3%	76.4%	66.2%

Tableau 2-1:étude comparative entre les travaux.

2.4 Conclusion :

D’après l’étude de ces 5 approches on a choisi l’approche d’extraction de concepts de texte basés sur WordNet arabe et analyse de concept formel car la ressource lexicale utilisée pour l'extraction de concepts est très importante ce qui nécessite de la tester.

L’objectif principal du travail qu’on a choisi est de combiner l'approche qui exploite la ressource lexicale AWN avec le formalisme mathématique de FCA afin d'accroître la capacité d'extraire un ensemble de concepts, réduits et pertinents, décrivant précisément le contenu sémantique d'un texte arabe.

3

Chapitre 03 :

Conception et Réalisation

3.1 Introduction :

Ce chapitre est consacré aux étapes fondamentales de la réalisation de notre travail qui est celui du développement d'un système d'extraction de concepts.

Lors de ce chapitre, nous allons donner des définitions des outils et formalisme utilisées pour les deux approches talque la première est l'extraction de concept a base de wordnet arabe notre système suit tout un processus qui commence par une étape de segmentation en passant par le prétraitement (normalisation , tokenization et suppression des mots vides) enfin la vérification d'existence dans l'ontologie wordnet arabe, et pour la deuxième la même chose pour les deux premières étapes ensuite le calcul des propriétés pour construire le contexte formel et enfin la liste des concepts formel .Après nous allons expliquer brièvement l'architecture et ses étapes de système pour les deux approches propose.

3.2 Architecture global du système :

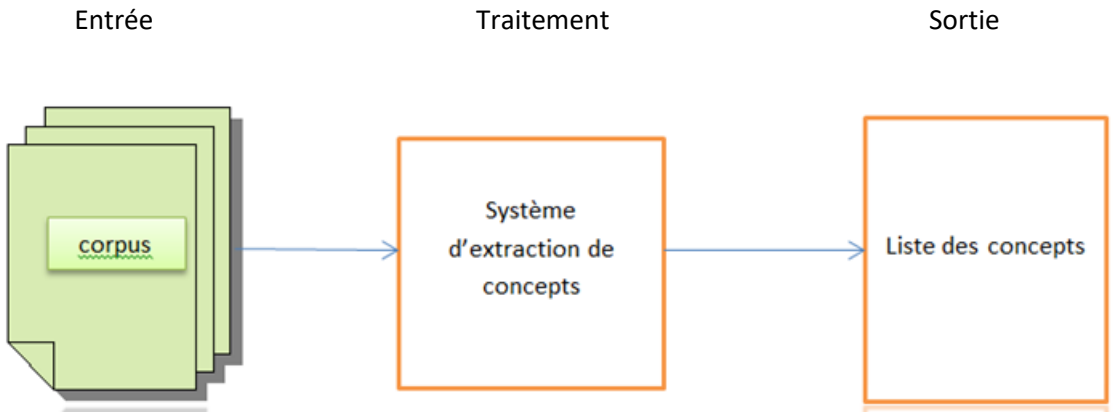

Figure 0-1: l'architecture globale de système d'extraction de concepts.

Description :

Entrée : est un texte écrit en arabe.

Système d'extraction de concepts : permet de faire des traitements sur le texte en entrée pour obtenir les concepts qui représentent la sémantique du texte.

Sortie : exprime le résultat de travail qui est une liste des concepts extraits à partir du système d'extraction

3.3 L'approche utilisée :

On a choisi l'approche présentée dans [10] « Text Concepts Extraction based on Arabic WordNet and Concept Formal Analysis» qui vise à extraire des concepts, représentant un texte arabe, ces concepts pouvant être utilisés plus tard pour la phase d'indexation. Mais on a choisi de faire une comparaison entre l'approche basée sur WordNet Arabe avec l'approche basée sur l'analyse de concept formel.

3.3.1 Extraction de concepts basée sur WordNet Arabe :

Cette approche est divisée en deux étapes :

La première étape : Extraction des termes

Cette étape consiste à extraire les termes qui représentent le texte. On segmente chaque mot dans le texte, ensuite normaliser le texte, puis tokeniser le texte et la fin de cette étape on va éliminer les mots vides pour construire la liste des mots.

La deuxième étape : Extraction des concepts

T = (T_1, T_2……T_n).Après nous projetons les termes de la liste produite sur AWN, afin de vérifier l'existence de chaque terme. Si ce terme existe dans AWN on l'ajoute dans la liste des concepts C = (C_1, C_2……C_n).

3.3.2 Extraction de concepts basée sur l'Analyse de Concept Formel(ACF) :

Cette approche est devisée en deux étapes :

Etape 01 :Nous commençons par extraire les trois expressions clés qui résument le contenu du texte a l'aide de l'algorithme Rapid Automatic Keyword Extraction (RAKE) qui est un algorithme d'extraction de mots clés indépendant du domaine qui tente de déterminer les expressions clés dans un corps de texte en analysant la fréquence d'apparition des mots et leur cooccurrence avec d'autres mots du texte. Cette étape étant effectuée avant le prétraitement du texte, de sorte que le sens ne soit pas perdu. Ensuite, nous allons effectuer un prétraitement de texte (comme la première approche) pour construire la liste des termes T = (T_1, T_2……T_n). Après nous allons calculer la similarité entre les termes obtenus avec les phrases des idées de texte.

Etape 02 : Dans le but d'obtenir des concepts, qui ont un degré important dans la représentation sémantique du texte. Dans cette étape, nous commençons par construire un contexte formel, à partir de la relation binaire (termes X (nombre d'occurrence et la similarité)). Plus particulièrement, nous considérons que, l'ensemble d'objets de ce contexte formel sont les termes identifiés lors de la première étape, et que l'ensemble de propriétés sont le nombre d'occurrence et la similarité par rapport aux idées de texte.

Nous construisons, par la suite, le treillis associé à ce contexte formel, et nous choisissons les rectangles maximal dont l'intention contient le plus grand nombre de propriétés. Nous générons, à partir de ce rectangle, les concepts contenus dans son extension, et nous obtenons ainsi, l'ensemble final de concepts.

3.4 Architecture de système:

3.4.1 Extraction de concepts basée sur WordNet Arabe :

Le processus de l'application de ce système suit une série de traitements, en commençant par un prétraitement des textes de corpus (normalisation, tokenization et élimination des mots vides) après un traitement sur la base lexicale WordNet (vérification et projection) afin de construire la liste de concepts.

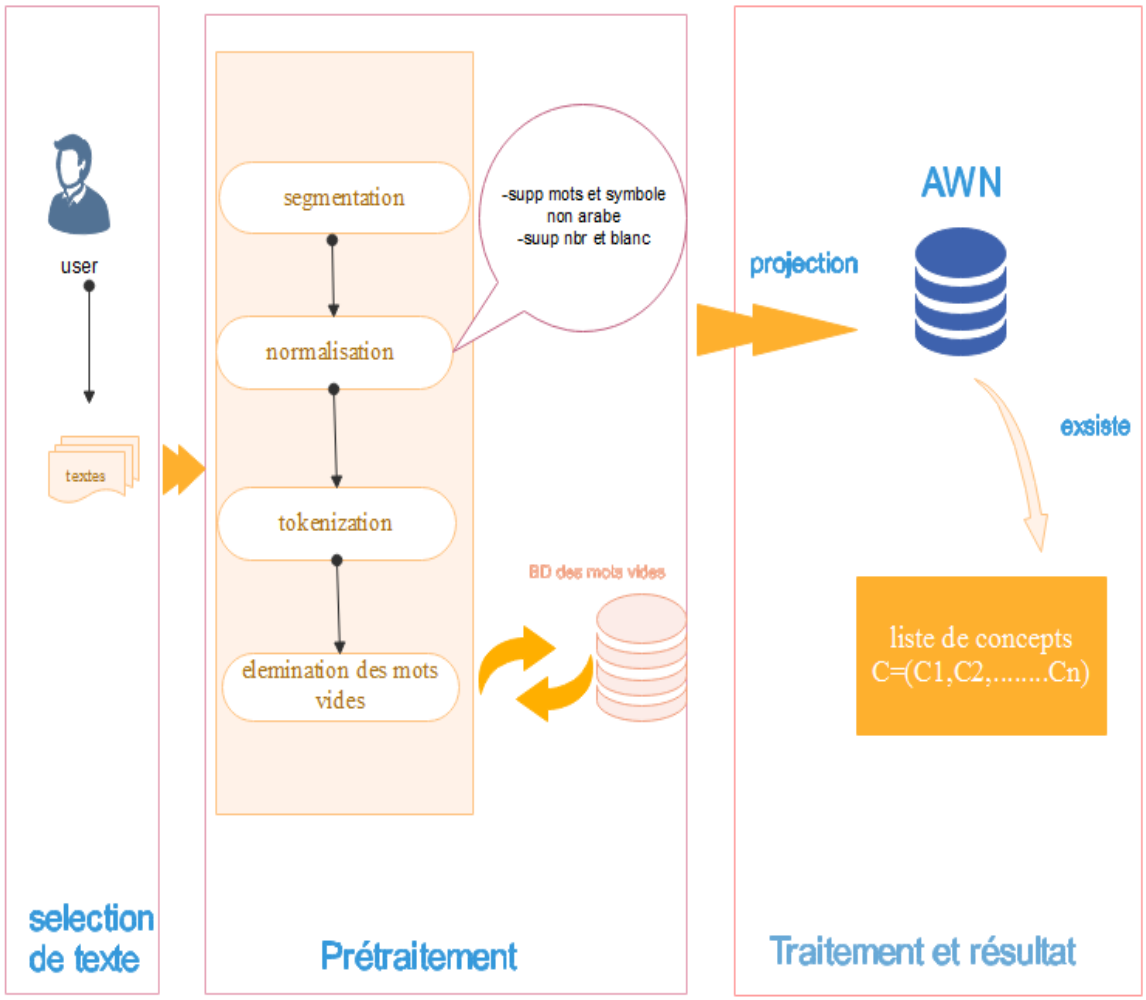

Figure 3-2:architecture détaillée de première approche.

3.4.2 Extraction de concepts basée sur l'Analyse de Concept Formel(ACF) :

Le processus de l'application de ce système suit une série de traitements, en commençant par un prétraitement des textes de corpus (normalisation, tokenization et élimination des mots vides). Ensuite, la construction de contexte formel et enfin construction du treillis Galois pour définir les concepts formels.

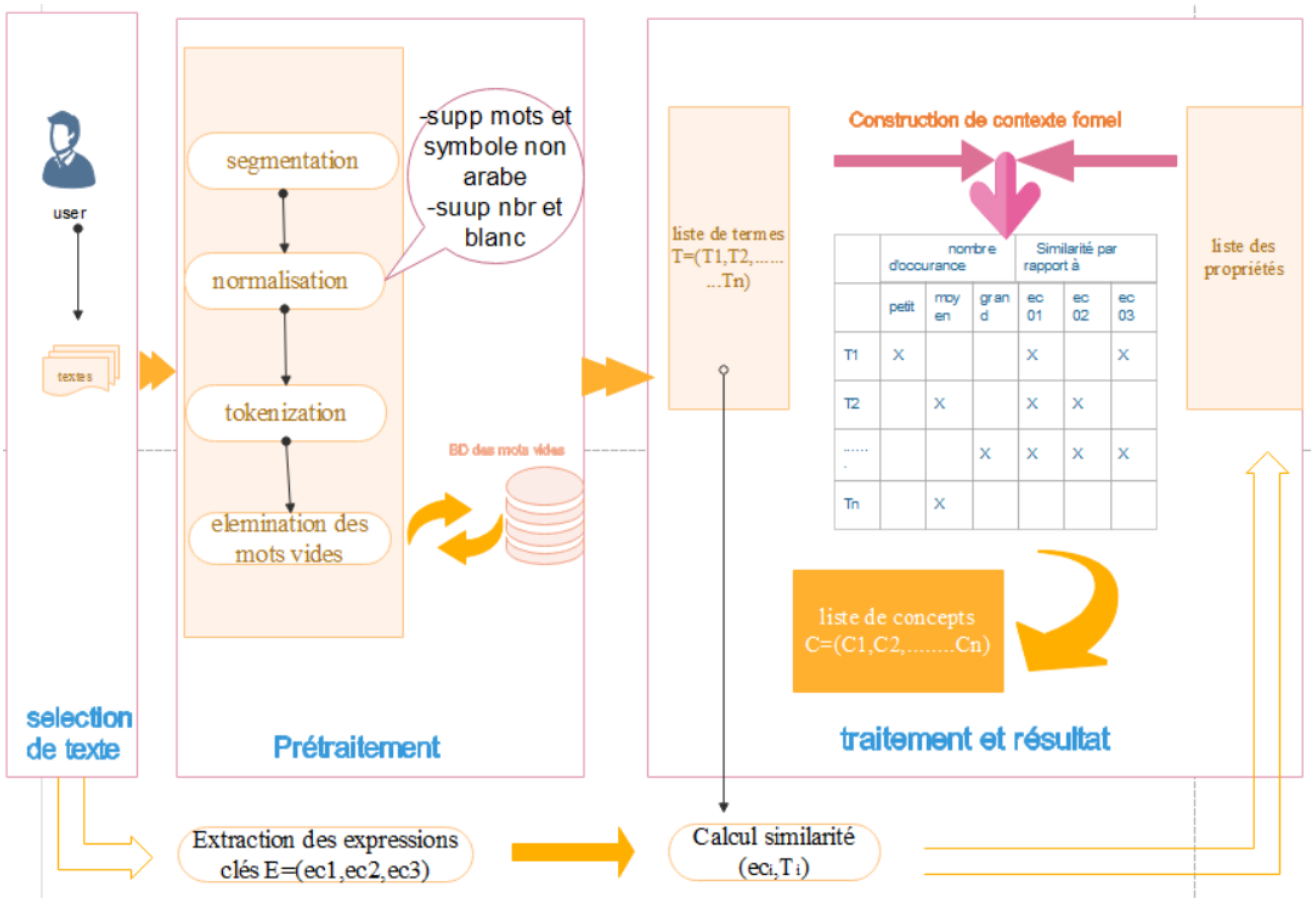

Figure 3-3:architecture détaillé de deuxième approche.

- **Prétraitement**

Segmentation :

La segmentation est une étape nécessaire et signifiante dans tout traitement de la langue naturelle [14]. La fonction d'un segmenter est de couper un terme courant en segments, de sorte qu'ils puissent être introduits dans un capteur morphologique ou dans un étiqueteur

de position pour un traitement ultérieur. Dans notre approche nous avons opté à une segmentation en tokens. Cette étape se faire avec « stanford segmenter ».

La normalisation :

Plusieurs genres de normalisation sur le texte sont appliqués afin de mieux manipuler les variations du texte qui peuvent être représentées en arabe. Par exemple, dans l'arabe écrit, les voyelles sont souvent omises dans les textes, néanmoins, on peut parfois trouver quelques voyelles présentées avec des mots. Alors, l'élimination de ces voyelles est nécessaire pour mettre fin à la normalisation. Certaines lettres subissent une simple modification dans l'écriture qui n'influe pas considérablement sur le sens du mot. Mais l'encodage de ces lettres change d'un mot à un autre. Une autre raison pour ce prétraitement est que l'on a tendance fréquemment à mal écrire ces différentes formes de hamza. Ce genre d'erreurs est très répandu dans les textes arabes. Par exemple, le mot « أكل « est généralement écrit « <اكل>.

La normalisation concerne les étapes suivantes :

- Suppression des mots et symboles non arabe.
- Suppression des chiffres
- Suppression des plusieurs blancs

Tokenization :

Est un processus de démarcation et éventuellement de classification des sections d'une chaîne de caractères en entrée. Les jetons résultants sont ensuite transmis à une autre forme de traitement. Le processus peut être considéré comme une sous-tâche de l'analyse des entrées.

Elimination des mots vide:

Les mots vides sont des mots non significatif figurant dans un texte dans la langue arabe il y a 750 mots vide[6]

Si mot existe dans liste des mots vide
Alors supprimer mot

Caractéristique de corpus :

Le corpus utilisé contient oui domaine, chaque domaine contient un ensemble de textes,

Comme le tableau suivant montre :

[6]https://github.com/mohataher/arabic-stop-words/blob/master/list.txt

Domaine du corpus	Nombre d'unité lexicale
(3102) اقتصاد	Min =123 Max =424
(3233) تاريخ	Min =269 Max =1754
(3608) تربية اسرة و مرأة	Min =231 Max =3130
(3171) دين وفتاوى شرعية	Min =90 Max =1761
(2419) رياضة	Min =194 Max =640
(2296) صحة	Min =153 Max =2656
(557) فلك	Min =57 Max =
(710) قانون	Min =153 Max =

Tableau 3-1:collection de test

Exemple de prétraitement d'une phrase:

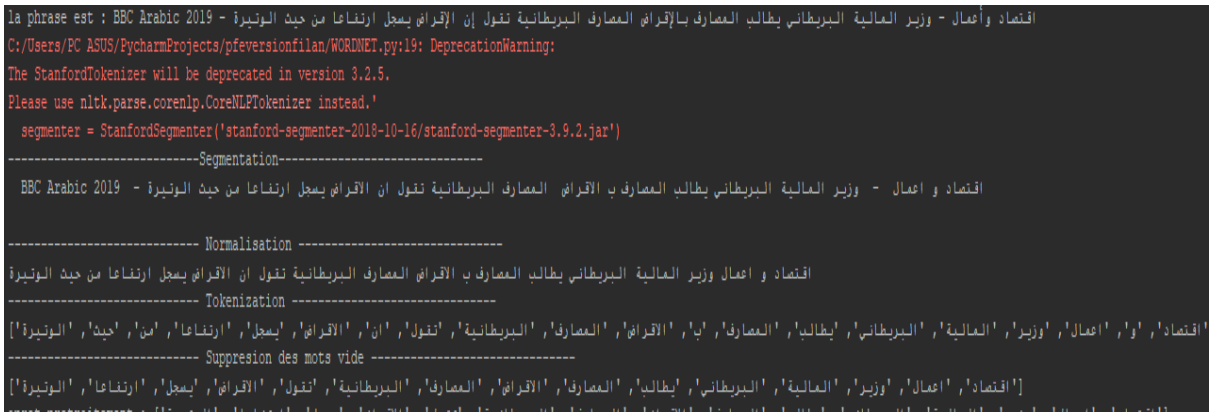

Figure 3-4 Exemple de prétraitement d'une phrase

3.5 Conclusion :

Dans ce chapitre, on a décrit notre système, dont l'objectif est de concevoir un système capable d'identifier automatiquement des concepts présents dans des textes arabes et plus précisément dans les revues électroniques. La conception sera mise en fonction dans le chapitre qui suit.

4

Chapitre 04 : Implémentation et Evaluation

4.1 Introduction:

Dans ce chapitre, nous traitons la mise en œuvre de notre système. Nous abordons les principaux outils et bibliothèques utilisés dans notre travail avec l'environnement de développement. Nous expliquons également le fonctionnement de système étape par étape.

4.2 Implémentation :

4.2.1 Outils et Environnement de développement :

Python[7] est le langage de programmation le plus utilisé dans le domaine du Machine Learning, du Big Data et de la Data Science.

Créé en 1991, le langage de programmation Python apparu à l'époque comme une façon d'automatiser les éléments les plus ennuyeux de l'écriture de scripts ou de réaliser rapidement des prototypes d'applications.

Depuis quelques années, toutefois, ce langage de programmation s'est hissé parmi les plus utilisés dans le domaine du développement de logiciels, de gestion d'infrastructure et d'analyse de données. Il s'agit d'un élément moteur de l'explosion du Big Data.

PyCharm[8] est un environnement de développement intégré utilisé pour programmer en Python.

Il permet l'analyse de code et contient un débogueur graphique. Il permet également la gestion des tests unitaires, l'intégration de logiciel de gestion de versions, et supporte le développement web avec Django.

Développé par l'entreprise tchèque JetBrains, c'est un logiciel multi-plateforme qui fonctionne sous Windows, Mac OS X et Linux. Il est décliné en édition professionnelle, diffusé sous licence propriétaire, et en édition communautaire diffusé sous licence Apache.

Sqlite Browser est un logiciel gratuit, domaine public, outil visuel open source utilisé pour créer, concevoir et éditer des fichiers de base de données compatible avec sqlite.

[7] https://www.lebigdata.fr/python-langage-definition
[8] https://fr.wikipedia.org/wiki/PyCharm

RazorSQL est un outil de requête SQL, un navigateur de base de données, un éditeur SQL et un outil d'administration de base de données pour Windows, macOS, Mac OS X, Linux et Solaris. RazorSQL a été testé sur plus de 40 bases de données, peut se connecter à des bases de données via JDBC ou ODBC.

4.2.2 Description :

Le langage de programmation qu'on a utilisé dans notre application est python 3.6 dans l'environnement PyCharme qui permet d'utiliser les différentes API (Application Programming Interfaces) à partir de plusieurs langages de programmation tels que Java, où on a utilisé certain fonctions de prétraitement du texte. Des fonctions d'accès à la base de données d'AWN et des fonctions de vérification.

On a exporté les tables de WordNet Arabe dans Sqlite3 pour faciliter l'utilisation de la base de données

Nous avons utilisé des fonctions fournies par différentes bibliothèques, nous citons:

Nltk (Natural languagetoolkit): est une bibliothèque logicielle en Python permettant un traitement automatique des langue développée par Steven Bird et Edward Loper du département d'informatique de l'Université de Pennsylvanie.

Stanford[9] : stanfordCoreNLP fournit un ensemble d'outils technologiques en langage humain. Il peut donner les formes de base des mots, leurs parties du discours, qu'il s'agisse de noms de sociétés, de personnes, etc., normaliser les dates, les heures et les quantités numériques, marquer la structure des phrases en termes de syntagmes et de dépendances syntaxiques, indiquer quelles expressions nominales font référence aux mêmes entités, indiquent un sentiment, extraient des relations particulières ou à classes ouvertes entre des mentions d'entités, obtiennent les citations que les gens ont dites, etc. Nous avons l'utiliser pour la segmentation des textes.

Pyarabic : une bibliothèque de langue arabe spécifique à Python, fournit des fonctions de base pour manipuler les lettres et le texte arabes, telles que la détection des lettres arabes, les groupes de lettres arabes et leurs caractéristiques, supprimer les signes diacritiques, etc.

PyQt est un module libre qui permet de lier le langage Python avec la bibliothèque Qt. Il permet ainsi de créer des interfaces graphiques en Python. Une extension de QtDesigner (utilitaire graphique de création d'interfaces Qt) permet de générer le code Python d'interfaces graphiques.

[9] https://stanfordnlp.github.io/CoreNLP/

Concepts[10] : est une simple implémentation Python de la ACF.

difflib.SequenceMatcher : Il s'agit d'une classe flexible permettant de comparer des paires de séquences de tout type, dans la mesure où les éléments de séquence sont hashable. L'algorithme de base est antérieur et un peu plus sophistiqué qu'un algorithme publié à la fin des années 1980 par Ratcliff et Obershelp sous le nom hyperbolique de «correspondance de modèle de gestalt». L'idée est de trouver la sous-séquence de correspondance contiguë la plus longue ne contenant aucun élément indésirable. (L'algorithme Ratcliff et Obershelp ne traite pas les pourriels). La même idée est ensuite appliquée récursivement aux parties des séquences situées à gauche et à droite de la sous-séquence correspondante

Rake : est un algorithme d'extraction de mot clé (Rapid Automatique Keyword Extraction) indépendant du domaine qui essaie de déterminer les phrases clés dans un corps de texte en analysant la fréquence d'apparition du mot et sa co-occurrence avec d'autres mots du texte.

4.2.3 Déroulement :

On présente les déférentes étapes de déroulement du processus d'extraction de notre système depuis le choix des textes jusqu'à l'identification des concepts. En commençant par l'interface principale de notre système dans la figure suivante :

Figure 4-1:Interface principale de système.

[10] https://concepts.readthedocs.io/en/stable/

L’interface de notre système est simple et claire. Elle contient des boutons et bondes de lecture et affichage .Elle fonctionne comme suit :

- ""اختيار ملفpour sélectionner un texte.si l’utilisateur veut saisir un texte il peut le saisir dans la bande "يرجى ادخال النص"
- ""النتيجة pour afficher le résultat dans la bonde "المفاهيم المستخرجة"
- Pour choisir la méthode il faut cliquer sur AWN ou ACF.
- ""معالجة النص pour choisir les étapes qu’il veut voir.
- "حذف" concerne le nettoyage des résultats précédentes.
- "خروج" exit l’interface du système/

Figure 4-2: Normalisation d’un texte

Figure 4-3: La segmentation d'un texte

Figure 4-4: Le résultat selon ACF

Figure 4-5: Le résultat selon AWN

4.2.4 Exemple de traitement d'un texte :

Texte :

موقع القانون الليبي-Law Of Libya -
قوانين بيئية تقبل تعيين خريج جامعة خاصة غير معتمدة عضو هيئة تدريس بالجامعة بمنصب إداري منصب أمين لجنة شعبية يتدرج فى السلم الاداري لاكتساب خبرة قوانين بيئية انعكاسات العولمة على اوضاع المراة العاملة بدعوة من مركز المرأة العربية كوثر شاركت د فائزة ببحث فى المؤتمر الذي عقد لدراسة انعكاسات العولمة على اوضاع المراة العاملة بتونس (0) المؤتمر المدني الموازي لمتندي المستقبل عمان بدعوة من مركز عدالة بالأدرن بالفترة من 27-28 / 11 / 2006م شاركت د فائزة بالمؤتمر المدني الموازي لمنتدي المستقبل عمان (0) المراة حقوق و واجبات بتاريخ ألقت د فائزة محاضرة بعنوان حقوق وواجبات المرأة استهدفت الموظفات بشركة البريقة طرابلس (0) الفساد الاداري واليات مكافحتة ألقت د فائزة الباشا محاضرة حول الفساد الاداري وآليات مكافحتة ولأول مرة يناقش هذا الموضوع المكان : مركز دراسات وأبحاث الكتاب الأخضر - طرابلس بتاريخ 2005/10/17م (0) العدالة التصالحية في المسائل الجنائية بتاريخ 2004/4/6 قدمت د. فائزة محاضرة بعنوان العدالة التصالحية في المسائل الجنائية (0) الرقابة الشعبية في ليبيا الرقابة الشعبية في ليبيا محاضرة ألقتها د فائزة المركز العالمي لدراسات وأبحاث الكتاب الأخضر بتاريخ 2004/12/27م (0) الرقابة الثورية ابعاد المهمة وكيفية التطوير بتاريخ 6 / 6 / 2005 قدمت محاضرة الرقابة الشعبية فىليبيارؤية تحليلية لندوة الرقابة الثورية ابعاد المهمة وكيفية التطوير تنظيم شعبة التثقيف والتعبئة بمثابة أبومليانة طرابلس (0) التشريعات الطبية بين الواقع والطموح برعاية من النقابة العامة للأطباء عقدت ندوة دولية بعنوان التشريعات الطبية بين الواقع والطموح قدمت د فائزة بحث بعنوان قراءة فى القانون رقم بشأن المسئولية الطبية (0) البرنامج الثقافي لكلية القانون البرنامج الثقافي لكلية القانون ناقش أعضاء هيئة التدريس مشروع قانون

العقوبات (0) اصدار تقرير عربي حول العنف ضد المراة اصدار تقرير عربي حول العنف ضد المراة الاردن (0) ...المزيد اجتماع الوفد النسائي اليوناني كيف يطلق سراح تجار المخدرات مبكرا لمادا عضو هيئة التدريس الليبي ضرورة تجاوز عقلية الاقصاء و التهميش دعوة لنبذ الشخصنة الي من يهمه امر الاجيال المرأة في حياة الكوني اللهم غير مفتونين القيم في عالم متغير العدالة البطيئة والموت البطي البداية السابق 11 12 13 14 15 التالى الأخير النتائج 241 - 260 من 281 [عودة] الرئيسية آخر الأخبار الإصدارات النشاطات إستشارات المقالات الأبحاث والدراسات مركز حقوق الانسان تشريعات تهمك قرارات تهمك المنتدى نشرة المنتدى غرف المحادثة روابط مفيدة أنشر الموقع إعلن في - - 2007 768 By الموقع إتصل بنا من نحن 1024

Le résultat :

- Les concepts extraits par ACF :

- ('المرأة',)
- ('المستقبل',)
- ('قوانين', 'الجامعة', 'العاملة', 'المستقبل', 'التشريعات', 'العامة', 'البرنامج')
- ('العربية', 'المؤتمر', 'الموظفات', 'الكتاب', 'المركز', 'العقوبات', 'الاصدارات', 'المقالات')

- Les concepts extraits par AWN :

موقع', 'تقبل', 'تعيين', 'جامعة', 'عضو', 'هيئة', 'تدريس', 'منصب', 'لجنة', 'شعبية', 'اكتساب', 'خبرة', 'دعوة', 'مركز', 'بحث', '] , 'عقد', 'دراسة', 'تونس', 'المدني', 'عمان', 'تاريخ', 'محاضرة', 'عنوان', 'شركة', 'طرابلس', 'مرة', 'الموضوع', 'المكان', 'ليبيا', 'ندوة', 'تنظيم', 'شعبة', 'رعاية', 'قراءة', 'رقم', 'كلية', 'ناقش', 'مشروع', 'قانون', 'اصدار', 'تقرير', 'عربي', 'الاردن', 'اجتماع', 'سراح', 'ضرورة', 'تجاوز', 'عقلية', 'نبذ', 'امر', 'حياة', 'عالم', 'متغير', 'الموت', 'عودة', 'تهم', 'غرف', 'اتصل

4.3 Evaluation :

Pour tester la performance de notre système d'extraction automatique de concepts on va faire d'abord un test manuel sur 10 textes en arabe (extraction manuelle de concept) ensuite, nous allons comparer le résultat du test manuel avec le test automatique par le calcule de rappel et précision selon les règles suivantes :

Le rappel (R) : est une évaluation de la couverture du système. Il mesure la quantité de réponses pertinentes d'un système par rapport au nombre de réponses idéales.

$$rappel = \frac{le\ nombre\ des\ concepts\ correct\ touvé}{le\ nombre\ total\ de\ concepts\ \ correcte} \qquad (1)$$

La précision (P) : est une évaluation du bruit du système. Elle mesure la proportion des réponses correctes parmi l'ensemble des réponses fournies par le système.

$$précision = \frac{le\ nombre\ de\ concepts\ correcte\ trouvé}{le\ nombre\ de\ concepts\ \ troué\ par\ le\ système} \qquad (2)$$

$$R = 1 - silence$$

$$P = 1 - bruit$$

Où :

- Silence : les concepts corrects non trouvé par le système.
- Bruit : les concepts trouvés qui normalement ne sont pas des concepts.

Après on va calculer la moyenne de rappel et précision entre les dix textes.

Les résultats des dix testes sont affichés dans le tableau suivant :

	Analyse de concepts formels		ArabicWordnet	
Teste	Rappel	Précision	Rappel	Précision
Test 01	0 .6	0.76	0.36	0.38
Test 02	0.6	0.6	0.3	0.3
Test03	0.7	0 .7	0.5	0.3
Test 04	0.2	0.5	0.1	0.03
Test 05	0.47	0.6	0.4	0.2
Test 06	0.65	0.9	0.3	0.3
Test07	0.7	0.57	0.35	0.33
Test 08	0.5	0.6	0.36	0.2
Test 09	0.5	0.4	0.2	0.1
Test 10	0.54	0.6	0.18	0.1
La moyenne	0.54	0.62	0.3	0 .24

Tableau 0-1: Les résultats rappel et précision des deux approche du système

Discussions :

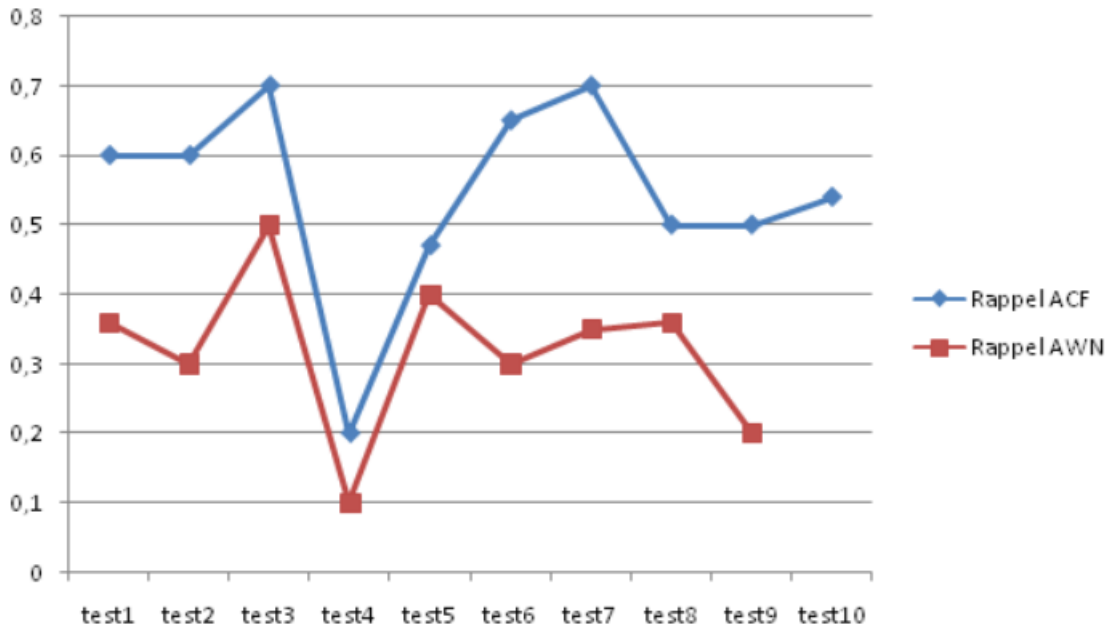

Figure 0-6: Graphe de rappel pour ACF et AWN

Nous remarquons que le traitement de l'ACF donne une moyenne de rappel par rapport au dix tests égale à 0.54. Cela signifie que le système peut extraire plus de la moitié des concepts du texte et reste silencieux sur 46%. Par contre AWN possède une moyenne de rappel égale à 0.3. C'est à dire que le système peut extraire 30% des concepts du texte et reste silencieux sur 70%.. Nous notons également que les deux courbes sont proportionnelles, ce qui signifie que le de texte joue un rôle important.

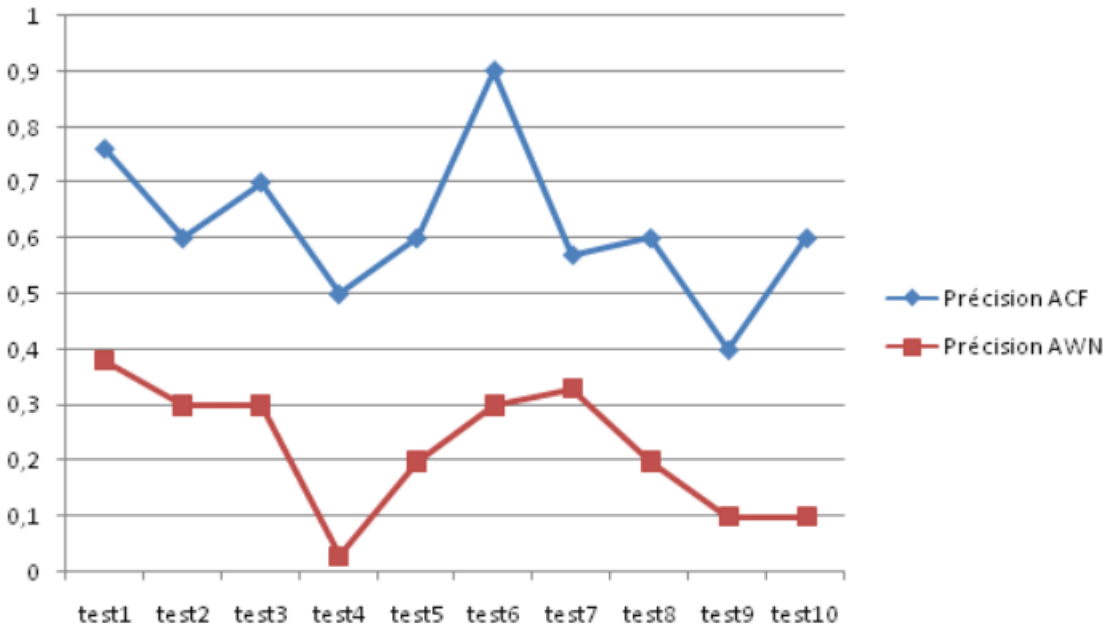

Figure 4-7:Graphe de précision de l'ACF et AWN

ACF a de précision 0.62, cela signifie qu'à partir du concept trouvé par le système 62 % sont corrects et par rapport à l'AWN précision = 0.24, c.-à-d. 24% de concepts sont correctes.

4.4 Conclusion :

A travers ce chapitre, nous avons présenté l'environnement de développement de notre système ainsi que l'interface graphique qui à travers laquelle nous pouvons superviser les différents traitements du système. Notre système a pour rôle d'extraire les concepts à partir des textes électroniques écrits en langue arabe. Les résultats de notre système sont liés étroitement au choix de la technique utilisée et à la qualité des textes traités.

Conclusion et perspectives :

Tout au long de ce mémoire, nous avons abordé la problématique de l'extraction de concepts à partir des documents textuels qui est un domaine très important dans le traitement automatique de a langue.

A travers notre projet, nous avons pu réaliser un système capable d'extraire automatiquement les concepts à partir des textes écrit en langue arabe. Nous avons étudié cette langue d'un point de vue informatique en faisant ressortir les traits linguistiques et statistiques à partir du texte.

Nous avant utilisé une méthode basé sur wordnet et l'analyse de concept formel.

La méthode que nous avons adoptée s'est avérée adaptable et nous avons pu faire nos expérimentations et évaluations sur un corpus qui regroupe un ensemble d'articles électroniques écrit en arabe.

Ce travail a permis d'acquérir nos connaissances dans le domaine de la programmation python, et de conforter nos connaissances en traitement linguistiques.

Comme perspectives, nous pensons que :

- l'ajout d'autres traits linguistiques comme la synonymie pour la détection et la sélection des concepts ayant le même sens améliorera énormément les résultats.
- la segmentation en tokens semble insuffisante et provoque des défauts de sens, par ailleurs, si on prend en considération les mots composés nous obtiendrons probablement de meilleurs résultats.
- L'enrichissement de WordNet arabe aussi peut faire différence car il est un des projets les plus importants dans le traitement automatique de la langue.

Bibliographie

[1] Khodja,AD. Un système d'extraction d'information pour la langue arabe.

[2] Barigou,F.(2013). Contribution à la catégorisation de textes et à l'extraction d'information (Doctoral dissertation).

[3] Dupont, M., Vuillaume, J. M., Victorri, B., Enjalbert, P., Mathet, Y., & Malandain, N. (2002). Nouvelles perspectives en extraction d'information.

[4] Bousmaha,Z.(2017).Une plateforme pour la conceptualisation des textes en langue arabe(Doctoral dissertation).

[5] Chouchaoui,M. & Brahimia,YA. (2016). Détection Automatique De La Cohésion Lexicale Entre Phrases Dans Les Textes Arabes.

[6] Bouhalassa, F., Mellal, N., & Guerram, T. (2018). Un outil d'extraction automatique de concepts à partir de donnée textuelles.

[7] Bouhriz, N., Benabbou, F., & Benlahmer, H. (May 2015). Une approche d'extraction basée sur la sémantique contenue dans un texte arabe.

[8] Assaghir, Z. (2010). Analyse formelle de concepts et fusion d'informations: application à l'estimation et au contrôle d'incertitude des indicateurs agri-environnementaux (Doctoral dissertation, Institut National Polytechnique de Lorraine).

[9] Karoui, L., Seghouani, N. B., & Aufaure, M. A. (2006, September). Extraction de concepts guidée par le contexte.

[10] Bouhriz, N., Benabbou, F., & Benlahmer, H. (2015). Text conceptsextraction based on Arabic wordnet and formal concept analysis. International Journal of Computer Applications, 111(16), 30-34.

[11] Lichouri, M., Djeradi, A., & Djeradi, R. (2015). Une approche Statistico-Linguistique pour l'extraction de concepts sémantiques: Une première étape vers un système générique de dialogue Homme-Machine. Personal Communication.

[12] Dlugolinský, Š. (2016). Combining Named Entity Recognition Methods for Concept Extraction. Information Sciences & Technologies: Bulletin of the ACM Slovakia, 8(2).

[13] Soldaini, L., & Goharian, N. (2016). Quickumls: a fast, unsupervised approach for medical concept extraction. In MedIR workshop, sigir.

[14] Boubekeur,Y. (2016). Identification automatique de mots clés dans les textes arabes.

[15] Harrathi, F. (2009). Extraction de concepts et de relations entre concepts à partir des documents multilingues: approche statistique et ontologique (Doctoral dissertation, Lyon, INSA).

[16] Even, F. (2005). Extraction d'Information et modélisation de connaissances à partir de Notes de Communication Orale (Doctoral dissertation).