الجـــــــــــــــــــــــمهـوريـــــة الجـــــــــــــزائـريـة الديمـــــــقـراطيـة الشـــــــــعبيـة

République Algérienne Démocratique et Populaire

وزارة التـــــــــــــــــــــــــــعـليـم العـــــــالي و البـــــــــحث العـــــلـمـــي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Centre Universitaire Salhi Ahmed – NAAMA
Institut des Sciences et de Technologie
Département de Mathématiques et Informatique

MEMOIRE

En vue de l'obtention du **diplôme** de **MASTER Académique**

En : INFORMATIQUE

Spécialité : **Système informatique**

Présenté Par : MAHAMMEDI BOUDJEMA

FENDOU MUSTAPHA

Intitulé

Classification des Questions

En langue arabe

Soutenu, devant le jury composé de :

Président			
Encadreur	Bouziane Abdelghani	MCB	INFORMATIQUE
Co-encadreur	/		
Examination	YAHIAOUI YASER	MCB	INFORMATIQUE
Examination	BENDIDA AISSAM	MAA	INFORMATIQUE

Session : (Mois juillet 2021)

Promotion : 2020 / 2021

REMERCIEMENTS

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 coeur 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 BOUZIANE Abdelghani** Docteur en science
à l'université de Naâma pour nous avoir encadrées et
Conseiller au cours de notre travail et
pour sa patience, sa disponibilité.
Nous souhaiterons également à remercier l'ensemble de
Nos professeur de l'université d'avoir nous accompagner durant le cursus universitaire et
les jurés d'avoir d'évaluer ce travail et pour leurs lectures attentives et leurs remarques
constructives.
En fin, nous remercions nos amies, nos collègues à toute la promo 2020/2021 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

Table des matières

Liste des figures

Liste des tableaux

Liste des Acronymes

QALD	Question Answering over Linked Data
NLIDB	Natural Language Interface to Data Bases
AQALD	Arabic Question Answering over Linked Data
SVM	Support Vector Machines
SPARQL	Simple Protocol and RDF Query Language
TALN	Traitement Automatique du Langage Naturel
TF-IDF	Terme Frequency Inverse Document Frequency
BOW	bag of Word

Introduction général

Aujourd'hui, la surcharge d'information est devenue de plus en plus un défi que les systèmes d'information doivent prendre en charge. En effet, nous remarquons l'expansion du contenu disponible sur différents médias (en particulier Internet). Par conséquent, il serait intéressant de mettre en place des outils permettant d'automatiser les traitements liés à la recherche de l'information, de faciliter l'accès à celle-ci, de diminuer la surcharge d'information, etc.

Jusqu'à aujourd'hui le marché de l'informatique essaie de répondre à cette problématique en développant des outils spécifiques tels que : les moteurs de recherche, les systèmes de Question/Réponse, les systèmes d'extraction d'information, les analyseurs morphologiques et syntaxiques, etc.

Notons que ces outils peuvent présenter une certaine forme de dépendance : par exemple, un système Q/R (ou d'extraction de l'information) peut faire appel à un analyseur morphologique. La maturité et l'efficacité de ces outils diffèrent selon le niveau de complexité du domaine traité et selon la langue cible. A ce titre, et malgré divers efforts, la maturité et l'efficacité de ce type d'outils pour le cas de la langue Arabe, reste relativement faible par rapport à d'autres langues.

La langue arabe fait partie de la famille des langues sémitiques et c'est la plus parlé avec près de 300 millions de locuteurs de la première langue.

La langue arabe a son propre script (écrit de droite à gauche) qui est une 28 lettre alphabet (25 consonnes et 3 voyelles longues) avec variantes allographiques et signes diacritiques qui sont utilisées comme voyelles courtes sauf un diacitique qui est utilisé comme double consonne marqueur.

L'écriture arabe ne prend pas en charge les majuscules. Les nombres sont écrits à partir de gauche à droite, ce qui constitue un véritable défi pour les éditeurs de texte arabe pour gérer les mots écrit de gauche à droite et d'autres de droite à gauche sur la même ligne.

La classification des questions est un élément très important du système question réponse. Dans ce mémoire, nous présentons une étude comparative de l'apprentissage automatique pour la classification des questions arabes. Nous avons utilisé deux

taxonomies : la taxonomie arabe et la taxonomie Li & Roth [21]. Nous avons mené plusieurs expériences sur nos questions du Jeu de données de la question arabe

Chapitre1

Système question réponse

1) Introduction

Les systèmes de question réponse doivent répondre à des questions plus précises que les systèmes de recherche d'information pour satisfaire au mieux les besoins des usagers. Un système de question réponse est une application qui cherche dans un corpus de document, une réponse exacte à une question posée en langue naturelle. En effet, face à une question telle que "Les étoiles se déplacent-elles dans le ciel ?", les moteurs de recherche traditionnels renvoient une liste de documents jugés pertinents par rapport à la question (figure 1), tandis qu'un système de question réponse, retourne à l'usager une réponse spécifique, telle que "oui". Les systèmes de question réponse qui existent actuellement ne sont pas orientés vers un domaine spécifique comme celui de l'apprentissage [12].

Figure 1:traitement automatique du Language naturel

Nous avons constaté que la majorité de ces derniers sont de nature indépendante et non intégrable aux systèmes d'apprentissage. D'une part, il n'y a pas de systèmes qui ont été développés dans l'esprit d'être réutilisés, modifiés, configurés selon les besoins des systèmes d'apprentissage. Par exemple, l'intégration ou le changement du corpus local de données demande le changement dans le code source du système pour tenir compte des nouvelles données. D'autre part, l'exactitude, la précision et la justification de la réponse sont des buts loin d'être réalisés. Le manque d'interaction avec l'usager et celui d'un corpus local de données sont des facteurs qui diminuent l'exactitude de la réponse. Actuellement, il n'existe, dans l'industrie du logiciel. Aucun système de question réponse qui soit orienté vers le domaine d'apprentissage.

2) Historique de SQR

Les premiers systèmes de traitement de la langue des années 1960-1970 étaient des systèmes de question-réponse sur des domaines restreints, que ces systèmes aient été développés dans un but de recherche d'information ou d'interface en langue naturelle. Ainsi, BASEBALL en 1963, SIR en 1968, LUNAR en 1973 et LADDER en 1977 pour une description de ces systèmes) permettaient tous trois d'interroger une base de connaissances structurée, en faisant éventuellement des inférences. BASEBALL disposait d'une base d'un an de faits concernant l'American League, et permettait de répondre à des questions du type «Combien de jeux ont joué les Yankees en juillet ?» ou «Contre qui ont perdu les Red Box le 5 juillet ?» ou encore «Est-ce que chaque équipe a joué au moins une fois dans chaque stade chaque mois ?». SIR répondait à propos de faits qu'il avait interprétés et stockés et enfin LUNAR et LADDER offraient une interface en langue naturelle sur une base de connaissances, représentant les informations sur la composition du sol lunaire rapportées par la mission Apollo 11 d'une part et des informations pour un système d'aide à la décision pour la Navy [17].

LUNAR [15] autorisait des questions enchaînées du type «Quels échantillons contiennent du P205 ? », Donne-moi les analyses du P205 dans ces échantillons». On peut toutefois noter des différences dans l'analyse des questions. Alors que BASEBALL fonctionnait par une décomposition en groupes et l'utilisation de mots-clés, SIR utilisait des patrons et LUNAR (ainsi que LIFER, l'interface de LADDER) intégrait une analyse de phrases par une grammaire ATN complétée d'une analyse sémantique.

Alors que pour tous ces systèmes le problème central n'était pas le traitement de questions, mais la recherche d'information ou la réalisation d'une interface en langue, Lehnert, avec son système QUALM, s'est posé le problème de la modélisation des questions et de l'élaboration de stratégies selon le type d'information demandé. Elle a ainsi proposé une classification en treize classes qui ont servi de base à bon nombre de systèmes actuels, même si certaines de ses catégories doivent être raffinées pour une recherche dans des textes portant sur tout domaine pour lesquels le processus d'interprétation est forcément moins approfondi. En effet QUALM permettait de poser des questions sur des histoires analysées par les systèmes SAM et PAM avec une représentation des informations sous forme de dépendances conceptuelles et de scripts. La stratégie de résolution des questions, et donc les heuristiques appliquées, différait selon la catégorie de

question. Cette catégorisation permettait aussi de déterminer le type d'information à fournir en réponse. Même si ce type d'approche peut théoriquement s'appliquer sur des textes portant sur tout domaine, elle requiert des bases de connaissances très structurées représentant la sémantique et la pragmatique de la langue et du monde en général, connaissances qu'il n'est pas envisageable ni même possible de fournir à un système. Néanmoins, une approche purement TAL peut être réalisée pour un domaine d'application limité, tel que cela a été fait dans le système Extrants qui répond à des questions portant sur les commandes Unix. Extrants repose principalement sur une analyse syntaxico-sémantique du manuel Unix permettant la réalisation d'inférences par la mise en œuvre d'un raisonnement logique. Le domaine étant celui des commandes disponibles dans un système informatique, on peut développer une base de connaissances précises pour le représenter, que ce soit pour le lexique, où on peut limiter les ambiguïtés, ou pour les connaissances sémantiques. Signalons que ce système propose aussi un mode de fonctionnement dégradé en cas d'échec de la première approche, mode reposant sur l'exploitation de mots-clefs.

Afin de passer outre la nécessité d'une analyse complète des textes pour répondre à des questions de tout type et portant sur tout domaine, le problème peut se poser différemment : au lieu d'envisager des questions dont les réponses sont des types très génériques, on précise le type d'information cherchée afin de la retrouver et l'extraire de textes. On se place ainsi dans le domaine de l'extraction d'information. Les conférences MUC4, organisée par le DARPA, qui ont vu le jour à la fin des années 80 et ont duré jusqu'en 96, ont permis une avancée rapide du domaine. A l'instar des systèmes de question-réponse, les systèmes en extraction d'information sont complexes et comportent de nombreux composants [23].

L'extraction d'information consiste à définir la requête par un patron, et à chercher à remplir ses différents composants selon l'information contenue dans les textes, le but était de retrouver des informations sur le terrorisme, à savoir celui qui a réalisé un attentat, quand, quelle victime, etc.

Les premiers systèmes visaient la réalisation d'une analyse complète des phrases des textes (des articles de journaux). Cette approche a été abandonnée au profit d'analyses de surface des textes, comme cela peut être illustré par les travaux du SRI qui a abandonné l'approche générique de TACITUS pour un système dédié, se focalisant sur l'information

cherchée uniquement. C'est ainsi que se sont notamment développées les recherches sur les entités nommées, à savoir le repérage de noms de personne, d'organisation, de lieu, de date, etc. et les systèmes ont réalisé des performances de plus de 90% dans cette tâche.

Les systèmes d'extraction d'information sont principalement fondés sur l'utilisation de patrons d'extraction, tels que <PERSON> was <killed/murdered> qui permet de remplir l'acteur de l'attentat par l'entité nommée PERSON. Ces patrons étaient principalement créés manuellement, même si des travaux concernant leur acquisition automatique, ou semi-automatique, sans qu'il y ait nécessité de disposer d'un gros corpus annoté ont vu le jour , mais sans apporter de véritable réponse au problème de devoir recréer, pour chaque type d'information cherchée, un ensemble de patrons d'extraction. Néanmoins, on peut considérer que ces systèmes ont jeté les bases de techniques reprises dans le cadre des systèmes de question-réponse [14].

Les différentes approches développées en extraction d'information sont-elles-aussi significatives de l'évolution du TAL en général, qui est passé de l'étude de mécanismes généraux jusqu'à la fin des années 80 et le début des années 90 à l'étude de mécanismes dédiés, notamment en analyse, avec l'étiquetage morpho-syntaxique, les analyses robustes de surface des phrases, le repérage de marqueurs dans des textes, les traitements statistiques, etc., approches qui ne requièrent pas nécessairement des bases de connaissances sémantiques.

L'idée de construire une large base de connaissances générales est cependant restée, et ont eu lieu notamment les travaux de Lenat sur la constitution de bases encyclopédiques, et les travaux sur Word Net, visant la construction d'une base lexicale sémantique. Si le projet CYC pas donné lieu pour le moment à des avancées significatives dans le domaine de la compréhension de la langue, Word Net est largement utilisée. Cette base organise les concepts en une hiérarchie et fournit un ensemble de relations pouvant être lexicales telles la synonymie, ou conceptuelles telles l'hyperonymie. Des définitions en langue sont en outre associées aux concepts. Cette ressource est largement utilisée dans les systèmes de question-réponse.

Avec le niveau de réalisation atteint par les travaux précédemment cités, ainsi que la multiplication des documents en ligne, qui peuvent être considérés comme une très vaste base de connaissances même si elle n'est pas structurée, la problématique question-réponse a pu à nouveau être abordée, mais avec une vision différente des premières approches. Ainsi, lors de la conférence TREC-8 en 1999, la première évaluation de systèmes de question-réponse a vu le jour et connaît un important succès depuis lors. Le principe est de répondre à des questions factuelles, questions de type encyclopédiques dont la réponse peut être formulée en peu de mots, sans limitation du sujet, par extraction d'un passage de texte. La base de documents est constituée d'articles de journaux.

Le fait de s'intéresser à des questions factuelles et encyclopédiques portant sur des événements dont on a rendu compte dans des textes, et non à des questions visant à connaître les tenants et les aboutissants d'une histoire, permet de mettre en œuvre des méthodes efficaces reposant sur des ressources et outils actuellement disponibles. Dans ce contexte, il s'agit de retrouver une réponse telle qu'elle figure au sein d'un ensemble de textes, ensemble le plus vaste possible car sa taille est en relation avec les chances de trouver une réponse à la question posée. Des méthodes de type recherche d'information sont utilisées pour sélectionner des passages intéressants dans un grand corpus. Cette première sélection permet ensuite d'appliquer des traitements, génériques ou dédiés à cette tâche, reposant souvent sur des approches issues du TAL, pour analyser plus en détail ces passages. C'est ainsi que les modules TAL utilisés permettent un typage de la réponse attendue lors de l'analyse des questions, la reconnaissance d'entités nommées correspondant aux types de réponse gérés, ainsi que la gestion de la variation linguistique entre la formulation de la question et la formulation de la réponse dans les documents.

3) état d'art du SQR

3.1) Les systèmes question réponse pour les données liées QALD :

L'objectif de QALD, est de permettre aux utilisateurs ordinaires de poser des questions en langage naturel, en utilisant leur propre terminologie, et de recevoir une réponse exacte, on exploitant les données liée.

Depuis la croissance constante du WS et l'émergence de la sémantique à grande échelle, la nécessité de créer des systèmes question réponse vers les données liées QALD basés sur les ontologies est devenue plus importante. Cette tendance a également été soutenue par des études d'utilisabilité (Kaufmann et Bernstein, 2007), qui montrent que les

utilisateurs occasionnels, généralement dépassés par la logique formelle du WS, préfèrent utiliser QALD pour interroger le WS [13]. Les QALD combinent plusieurs sources de données structurées pour produire une réponse exacte a une question posé en langages naturel (Figure 2).

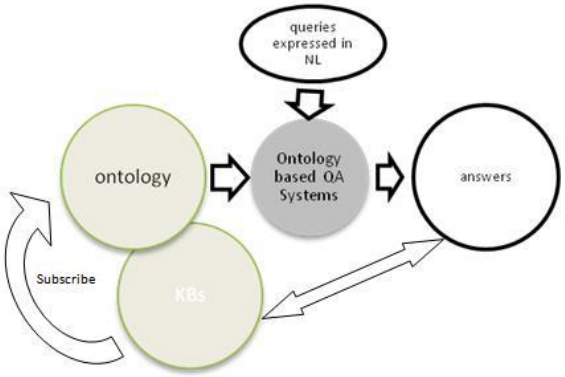

Figure 2:SQR basé sur les ontologies, entrée/sortie.

Ces dernières années, les QALD ont attiré beaucoup d'intérêt, où la puissance des données liée en tant que modèle de connaissance est directement exploitée pour l'analyse et la traduction des requêtes, offrant ainsi une nouvelle tournure pour l' ancienne génération des ILNBD, en se concentrant sur la portabilité et la performance, et en remplaçant les techniques couteuse du TAL utilisés pour domaine spécifiques par des techniques peu profondes mais efficace pour le domaine indépendant .

Les systèmes QR basés sur l'ontologie varient selon deux aspects principaux:

1. Le degré de personnalisation du domaine dont ils ont besoin, qui est en corrélation avec leur performance.

2. Le sous-ensemble de langage naturel qu'ils sont capables de comprendre (langage naturel d'une grammaire complète, langage naturel contrôlée ou guidée, langage naturel basée sur des patterns).

Afin de réduire à la fois la complexité et le problème d'habitabilité, qui sont les principaux problèmes qui entravent l'utilisation réussite des interfaces de langage naturel (Kaufmann and Bernstein 2007) [16].

3.1. Architecture globale :

Les QALD prennent en entrée les requêtes exprimées en langage naturel et une ontologie donnée et renvoient des réponses tirées d'une ou de plusieurs base de connaissance. Par conséquent, QALD n'exigent pas que l'utilisateur connaît le vocabulaire ou la structure de l'ontologie. Il existe beaucoup de travaux sur les QALD la majorité entre eux respect l'architecture globale basé sur trois module : l'analyseur de question, formulation la requête SPARQL et le générateur de réponse. (Figure 3)

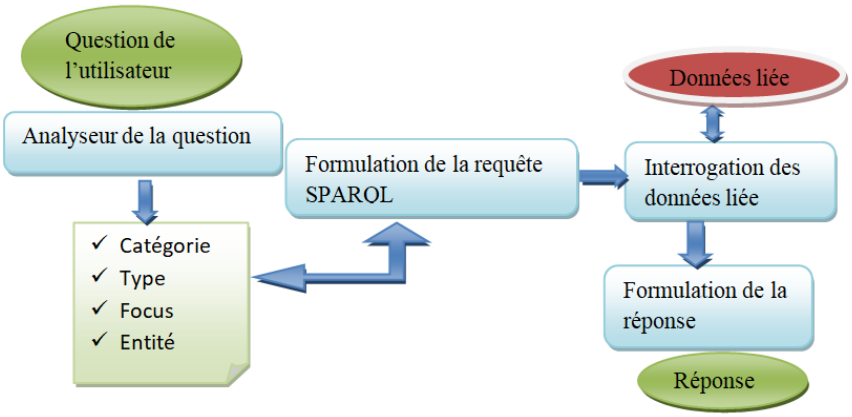

Figure 3:architecture globale des SQRDL

Pour SQR dédié au Web de données, l'Utilisateur pose une question en langage naturel. Le processus commence par une analyse de la question afin de déterminer les informations pertinentes telles que la catégorie de la question le type de la réponse, le focus, les entités nommées. L'étape suivante consiste utilisé les informations fournis par l'étape précédente afin de générer la requête SPARQL, Une ressource d'ontologie peut être utilisée pour enrichir ou faire la correspondance des éléments dans le processus. Enfin, lorsque la requête SPARQL est générée, l'interrogation des données liées est effectuée le résultat est traité pour générer la réponse exacte de la question de l'utilisateur.

3.1.1. Questions en langage naturel et les données liées :

La principale tâche des SQR est d'interpréter les informations de l'utilisateur exprimé en langage naturel par rapport aux données qui sont interrogées. Considérons un exemple simple: En ce qui concerne DBpedia, la question peut être exprimée au moyen de la requête SPARQL.

1. (a) Quelle est la monnaie de la République tchèque?

(b) SELECT DISTINCT?

Uri WHERE {res: Czech_Republic dbo: devise? Uri.}

Pour passer de la question à la requête, nous devons savoir que le nom la République tchèque correspond à la ressource res: Czech_Republic, que la devise d'expression correspond à la propriété dbo: monnaie, et nous devons connaître la structure de la requête, c'est-à-dire que l'entité res: Czech_Republic est la sujet de la propriété et que l'objet doit être retourné comme réponse.

Lors de la construction de la requête SPARQL à partir de la question est (relativement) simple dans cet exemple particulier, très souvent, le processus est beaucoup plus compliqué. Dans la plupart des cas, cela implique deux défis: la correspondance des expressions du langage naturel aux éléments de vocabulaire utilisés par les données, en tenant compte des discordances lexicales et structurelles, et en manipulant les variations de sens introduites par des expressions ambiguës et vagues, des expressions anaphoriques, etc. Regardons les deux défis suivants.

Mapper les expressions de langage naturel aux éléments de vocabulaire :

Les URI sont des identifiants indépendants de la langue. Bien qu'ils portent habituellement des noms mnémoniques, leur seule connexion réelle au langage naturel est par les étiquettes qui leur sont attachées. Ces étiquettes fournissent souvent une manière canonique de se référer à l'URI, mais ne tiennent généralement pas compte de la variation lexicale. La classe dbo: Film, par exemple, a le label anglais "Film" mais ne capture pas d'autres variantes telles que le Movie.

De même, la propriété dbo:spouse porte le label anglais spouse, tandis que la langue naturelle connaît une grande variété de façons d'exprimer cette relation, parmi lesquelles épouse, époux et marié avec , qui sont plus susceptibles de se produire dans la question de l'utilisateur que le terme un peu plus formel spouse. Tableau X présente les différences lexicale et les différences structurelle entre les questions est les requêtes.

3.1.2. Variantes de signification

Les SQR impliquent des processus de langage naturel et héritent ainsi des défis liés au traitement du langage naturel en général. L'un de ces défis concerne les ambiguïtés. L'ambiguïté couvre tous les cas où une expression en langage naturel peut avoir plus d'une signification, dans notre cas, elle peut correspondre à plus d'un élément de vocabulaire dans l'ensemble de données cible.

3.2 Natural Language interface to data bases (NLIDB)

Natural Language Interfaces est un domaine de recherche long. Le but de l'interface du langage naturel à la base de données Le système est d'accepter les demandes en anglais ou tout autre naturel la langue et tente de les « comprendre » ou nous pouvons dire que les interfaces de langage naturel avec les bases de données (NLIDB) sont systèmes qui traduisent une phrase en langage naturel dans un base de données. Bien que la recherche la plus commencé depuis la fin des années 1960, la NLIDB demeure ouverte problème de recherche.

Un système complet de la NLIDB nous sera utile à bien des égards. N'importe qui peut recueillir de l'information à partir de la base de données en utilisant de tels systèmes. En outre, il peut changer notre perception sur l'information dans une base de données. Traditionnellement, les gens sont habitués à travailler avec un formulaire; leurs attentes dépendent fortement sur les capacités de la forme [9].

NLIDB rend l'approche globale plus souple, ce qui permettra d'utilisation d'une base de données (figure 4).

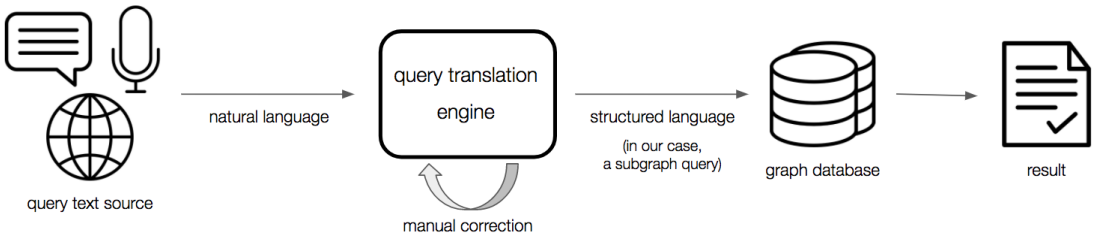

Figure 4:Interface en langage naturel vers les bases de données.

Il existe de nombreuses applications qui peuvent prendre des avantages de NLIDB. Dans les environnements de PDA et de téléphone portable, l'affichage écran n'est pas aussi large qu'un ordinateur ou un ordinateur portable. Remplir un formulaire qui a de nombreux champs peut être fastidieux : on peut avoir à naviguer à travers l'écran, pour faire défiler pour rechercher le défilement valeurs des cases, etc. Au lieu de cela, avec NLIDB, le seul travail qui doit être fait est de taper la question semblable au SGS (Système de messagerie courte).

3.2.1. Sous-composantes de la NLIDB

Les scientifiques de l'informatique ont divisé le problème de l'accès linguistique à une base de données en deux sous-composantes :

a - Composante linguistique

Il est chargé de traduire les entrées en langage naturel dans une requête formelle et génération d'une réponse en langage naturel en fonction des résultats de la recherche dans la base de données.

b- Composante de la base de données

Il exécute les fonctions traditionnelles de gestion des bases de données. A lexique est un tableau qui est utilisé pour cartographier les mots de la nature entrée sur les objets formels (noms de relation, noms d'attribut, etc.) de la base de données. Analyseur et interpréteur sémantique utiliser le lexique. Un générateur de langage naturel prend la réponse formelle comme son entrée, et inspecte l'arbre d'analyse dans afin de générer une réponse adéquate en langage naturel. Natural les systèmes de bases de données linguistiques utilisent les connaissances syntaxiques et la connaissance de la base de données réelle afin de faire le lien entre l'entrée du langage naturel et la structure et le contenu de cette base de données. La connaissance syntaxique réside habituellement dans le composante linguistique du système, en particulier dans la syntaxe analyseur alors que la connaissance de la base de données réelle réside dans une certaine mesure dans le modèle de données sémantiques utilisé. Questions entrée en langue naturelle traduite dans une déclaration dans un langage de requête formel. Une fois l'instruction sans ambiguïté formé, la requête est traitée par la gestion de la base de données afin de produire les données requises. Ces données passées à la composante du langage naturel où Les routines de génération produisent une version en langage de surface de la réponse.

3.2.2. Avantages et inconvénients de la NLIDB

Cette section traite des avantages et des inconvénients de NLIDB.

1. Aucun langage artificiel L'un des avantages des NLIDB est que l'utilisateur n'est pas nécessaire d'apprendre un langage de communication artificiel.

2. Simple, facile à utiliser Envisager une base de données avec un langage de requête ou un certain formulaire conçu pour afficher la requête. Alors qu'un système NLIDB ne nécessite qu'une seule entrée, un formulaire peut contenir entrées multiples (champs, boîtes de défilement, boîtes combinées, radio boutons, etc.) en fonction de la capacité du formulaire. Dans le dans le cas d'un langage de requête, une question peut devoir être exprimés à l'aide de plusieurs énoncés qui en contiennent un ou plus de sous-requêtes avec certaines opérations conjointes comme le connecteur.

3. C'est mieux pour certaines questions On a fait valoir qu'il y a un certain type de questions (p. ex. questions impliquant la négation, ou la quantification) qui peuvent être facilement exprimé en langage naturel, mais qui semble difficile (ou au moins fastidieux) à exprimer à l'aide de graphiques ou de formulaires Par exemple, « Quel ministère n'a pas programmeurs? » (Négation) ou « Quelle entreprise fournit chaque ministère? » (Quantification universelle), peut facilement être exprimé en langage naturel, mais ils seraient difficiles à s'exprimer dans la plupart des interfaces graphiques ou de forme. Questions comme le ci-dessus peut, bien sûr, être exprimé dans la requête de base de données langages comme SQL, mais langage complexe de requête de base de données des expressions peuvent devoir être écrites.

4. Tolérance aux pannes La plupart des systèmes de la NLIDB offrent certaines tolérances aux erreurs grammaticales, dans un système informatique; la plupart des temps, le lexique doit être exactement le même que défini, le la syntaxe doit suivre correctement certaines règles, et toutes les erreurs entraînera le rejet automatique de l'entrée par le système. Dans le cas de phrases incomplètes, la plupart des ordinateurs les systèmes ne fournissent aucun support.

5. Facile à utiliser pour plusieurs tableaux de base de données Les requêtes qui impliquent plusieurs tables de base de données comme «lister l'adresse des agriculteurs qui ont reçu une prime supérieure a à 10000 roupies pour la récolte de blé », sont difficiles à former sous forme graphique interface utilisateur par rapport à l'interface du langage naturel.

b. Inconvénients de la NLIDB

1. La couverture linguistique n'est pas évidente

À l'heure actuelle, tous les systèmes de la NLIDB ne peuvent traiter que certains sous-ensembles d'un langage naturel et il n'est pas facile de définir ces sous-ensembles. Même certains systèmes de la NLIDB ne peuvent répondre à certaines questions appartiennent à leurs propres sous-ensembles. Ce n'est pas le cas dans langue officielle. La couverture linguistique officielle est évidente. et toutes les déclarations qui suivent les règles données sont garanties pour donner la réponse correspondante.

2.Échecs linguistiques ou conceptuels

Dans le cas des défaillances du système de la NLIDB, il arrive souvent que le système ne fournit aucune explication des causes système à échouer. Certains utilisateurs peuvent essayer de reformuler le question ou tout simplement laisser la question sans réponse. La plupart des temps, il appartient aux utilisateurs de déterminer les causes de l'erreur.

3. Fausses attentes

Les gens peuvent être induits en erreur par la capacité d'un système du processus d'un langage naturel : ils peuvent supposer que le système est intelligent. Par conséquent, plutôt que de demander questions d'une base de données, ils peuvent être tentés de poser les questions qui impliquent des idées complexes, certains jugements, capacités de raisonnement, etc., qu'un système de la NLIDB ne peut s'appuyer sur.

3.2.3 Historique de la NLIDB

Les toutes premières tentatives d'interface de base de données NLIDB sont tout aussi anciennes comme toute autre recherche du PNL.

Poser des questions aux bases de données dans le langage naturel est très pratique et facile méthode d'accès aux données, spécialement pour les utilisateurs occasionnels qui ne comprennent pas langage complexe de requête de base de données tel que SQL.

Voici quelques exemples de l'interface du langage naturel pour Systèmes de base de données :

a-Système LUNAR

Est un système qui répond aux questions sur des échantillons de roches rapportées de la lune. Le système a été officieusement introduit en 1971. Pour accomplir son le système LUNAR utilise deux bases de données ; une pour analyse chimique et l'autre pour des références bibliographiques. Le système LUNAR utilise un réseau de transition augmenté
(ATN) et la sémantique procédurale de Woods. La performance du système LUNAR était tout à fait impressionnant; il a réussi à traiter 78% des demandes sans aucunes erreurs et ce ratio est monté à 90% lorsque les erreurs de dictionnaire étaient corrigées. Mais ces chiffres peuvent être trompeurs parce que le système n'a pas fait l'objet d'une utilisation intensive en raison de la limitation de ses capacités linguistiques.

b-LADDER

Le système LADDER a été conçu comme un langage naturel interface à une base de données d'informations sur les navires de la marine américaine. Le système LADDER utilise la sémantique grammaire pour analyser des questions pour interroger une base de données distribuée. Le système utilise une technique de grammaire sémantique qui s'entrelace traitement syntaxique et sémantique. La question répondre se fait par l'analyse de l'entrée et la cartographie de l'analyse à une requête de base de données [14].

Le système LADDER est basé sur trois couches d'architecture. Le premier composant du système est destiné à l'accès informel en langage naturel aux données de la Marine (INLAND), qui accepte les questions dans une langue naturelle et génère une requête vers la base de données. Les requêtes INLAND sont dirigés vers l'Intelligent Data Access (IDA), qui est le deuxième composant de LADDER.

Le composant INLAND construit un fragment d'une requête pour IDA pour chaque unité syntaxique de niveau inférieur en anglais requête d'entrée de langue et ces fragments sont ensuite combinés à des unités syntaxiques de plus haut niveau à reconnaître. Au niveau de la phrase, les fragments combinés sont envoyés sous forme de commande à l'IDA. L'IDA composerait une réponse pertinente à la requête originale de l'utilisateur en plus de planifier la bonne séquence de requêtes de fichiers. Le troisième composant du système LADDER est destiné au File Access Manager (FAM). La tâche de FAM est de

trouver l'emplacement des fichiers génériques et de gérer leur accès dans la base de données distribuée. Le système LADDER a été mis en œuvre dans le le système LADDER a permis de traiter base de données équivalente à une base de données relationnelle avec 14 tableaux et 100 attributs.

c- Système RENDEZVOUS

Ce système est apparu à la fin des années 70. Dans ce accéder aux bases de données par un langage naturel relativement libre.

Dans ce système Codd, l'accent est mis sur la requête paraphraser et engager les utilisateurs dans des dialogues de clarification lorsqu'il est difficile d'analyser les données de l'utilisateur.

d-PLANS

Cela a été développé à la fin des années 70 pour (Programmed Lnguage-based Enquiry System) à l'Université de l'Illinois Laboratoire scientifique coordonné. PLANES comprend un anglais langue avant-gardiste avec la capacité de comprendre et explicitement répondre aux demandes des utilisateurs. Il effectue des dialogues de clarification avec l'utilisateur ainsi que de répondre à des questions vagues ou mal définies. Ce travail est effectué à l'aide d'une base de données 3-M de la marine américaine (maintenance et matériaux gestion), il s'agit d'une base de données sur l'entretien des aéronefs et données de vol, bien que les idées peuvent être directement appliquées d'autres bases de données non hiérarchiques fondées sur les enregistrements.

e- PHILIQA

Il a été développé en 1977 et était connu sous le nom de Philips Système de réponse aux questions, utilise un analyseur syntaxique qui fonctionne comme un passage séparé de la compréhension sémantique passe. Ce système est principalement impliqué avec des problèmes de sémantique et a trois couches séparées de sémantique compréhension. Les calques sont appelés « anglais formel Language", "World Model Language" et "Data Base" Langue" et semblent correspondre approximativement au "externe", les vues "conceptuelles" et "internes" des données.

f- CHAT-80

Le système CHAT-80 est l'un des NLIDB les plus référencés dans les années 1980. Le système a été mis en Prolog. Le CHAT-80 était impressionnant, système efficace et sophistiqué. La base de données de CHAT-80 se compose de faits (c.-à-d. les océans, les grandes mers, les principaux cours d'eau et grandes villes) environ 150 pays du monde et un petit ensemble du vocabulaire de la langue anglaise qui suffisent pour interroger la base de données. Le système CHAT-80 traite une anglaise question sur la langue en trois étapes [6].

g-TEAM

Il a été élaboré en 1987. Une grande partie du temps a été consacré aux questions de transférabilité. TEAM a été conçu pour facilement configurable par les administrateurs de base de données sans la connaissance des NLIDB.

h-ASK

Ce système, mis au point en 1983, a permis aux système de nouveaux mots et concepts à tout moment au cours de la interaction. ASK était en fait une information complète système de gestion, fournissant sa propre base de données intégrée et la capacité d'interagir avec plusieurs bases de données externes, programmes de courrier électronique et autres applications informatiques. Toutes les applications connectées à ASK étaient accessibles à l'utilisateur demandes en langage naturel. L'utilisateur a déclaré demandes en anglais et Ask générées de manière transparente demandes aux systèmes sous-jacents appropriés.

i-JANUS

Il avait des capacités similaires pour interagir avec plusieurs sous-jacents systèmes (bases de données, systèmes experts, dispositifs graphiques, etc.).

Les systèmes sous-jacents pourraient participer à l'évaluation d'une demande en langage naturel, sans que l'utilisateur ne devienne conscient de l'hétérogénéité du système global. JANUS est aussi l'un des rares systèmes pour soutenir les questions temporelles.

j-EUFID

Le système EUFID se compose de trois modules principaux, non compter le SGBD. Le premier est le module analyseur, le second est module mapper et troisième module traducteur.

k-DATALOG

C'est un système de requête de base de données en anglais basé sur Cascade grammaire ATN. En fournissant des schémas de représentation séparés pour la connaissance linguistique, la connaissance du monde en général, et domaine d'application, DATALOG atteint un degré de portabilité et d'extensibilité. Systèmes qui apparurent au milieu des années 80 étaient LDC, TQA, TELI et bien d'autres.

3.2.4 Faits nouveaux récents concernant la NLIDB

Cette section donne un bref aperçu de trois Les systèmes du NLIDB ont récemment été mis au point dans différentes universités.

a. NALIX

NALIX (interface en langage naturel pour une base de données XML) est un système développé par le NLIDB à l'Université de Michigan, Ann Arbor de Yunyao Li, Huahai Yang et H. V. Jagadish (2006). La base de données utilisée pour ce système est base de données de langage de balisage extensible (XML) avec Schema-Gratuit XQuery comme langage de requête de base de données.

Schema-Free XQuery est un langage de requête conçu principalement pour récupérer des informations en XML. L'idée est d'utiliser recherche par mots clés pour les bases de données. Cependant, mot clé pur recherche ne peut certainement pas être appliquée. Par conséquent, certains plus riches des mécanismes de requête sont ajoutés. Compte tenu d'un ensemble de mots clés, chaque mot clé a plusieurs candidats XML éléments à relier. Tous ces candidats sont ajoutés à MQF (Mise au point significative de la requête) qui trouvera automatiquement toutes les relations entre ces éléments.

L'avantage principal de Schema-Free Xquery est qu'il n'est pas nécessaire de mapper une requête dans le schéma exact de la base de données, car il sera automatiquement trouver toutes les relations données certains mots clés.

NALIX peut être classé comme un système basé sur la syntaxe, puisque Les processus de transformation se déroulent en trois étapes : un arbre d'analyse, validant l'arbre d'analyse et traduisant l'arbre de parse à une expression XQuery. Cependant, comme implicite dans le papier, NALIX est différent du général approches basées sur la syntaxe; dans la façon dont le système a été construit: NALIX met en œuvre une technique d'ingénierie inversée par construire le système à partir d'un langage de requête vers le phrases.

b. PRECISE

PRECISE est un système développé à l'Université de Washington par Ana-Maria Popescu, Alex Armanasu, Oren Etzioni, David Ko et Alexander Yates (2004). The Target base de données est sous la forme d'une base de données relationnelle utilisant SQL comme le langage de requête. Il introduit l'idée de sémantique phrases traçables qui sont des phrases qui peuvent être traduit en une interprétation sémantique unique en analysant certains lexiques et contraintes sémantiques.

PRECISE a été évalué sur deux domaines de base de données. Le premier l'un est le domaine ATIS, qui consiste en des questions orales sur le transport aérien, leurs formulaires écrits et leur traductions en langage de requête SQL. Dans le domaine ATIS, 95,8% des questions ont été faciles à précision de 94 %.

Le deuxième domaine est le domaine GEOQUERY. Ce domaine contient information sur la géographie des États-Unis. 77,5 % des questions dans GEOQUERY sont sémantiquement tractable. En utilisant ces Les questions donnent une précision de 100 %.

c. WASP

L'analyse sémantique basée sur l'alignement de mots (WASP) est un système développé à l'Université du Texas, Austin par Yuk Wah Wong. Bien que le système soit conçu pour l'objectif plus large de la construction « d'un représentation symbolique et significative d'un langage naturel« phrase », il peut également être appliqué au domaine de la BNIDB. A logique prédicat (Prolog) a été utilisé comme la requête formelle langue.

WASP apprend à construire un parseur sémantique à partir d'un corpus de phrases en langage naturel annotées avec leurs correctes langues de requête formelle. Il ne requiert

aucune connaissance préalable de la syntaxe, parce que tout le processus d'apprentissage se fait en utilisant techniques statistiques de traduction automatique.

4) Conclusion

La recherche sur les Interfaces linguistiques. Avec l'avancement dans la matérielle puissance de traitement, de nombreux NLIDB mentionnés dans résultats prometteurs. Bien que plusieurs systèmes ont également été développés à ce jour pour une utilisation commerciale mais l'utilisation des systèmes de la NLIDB n'est pas généralisée et n'est pas une option standard pour l'interfaçage avec une base de données. Ce manque de l'acceptation est principalement attribuable au grand nombre de lacunes dans le système de la NLIDB afin de comprendre une langue.

Dans ce chapitre on a présenté un état d'art sur les systèmes question réponse, des systèmes capables de répondre à des questions exprimé en langage naturelle. La première génération a connu le développement des interfaces de langage naturel vers les BDD, la génération suivante les systèmes question réponse pour les donnée textuelle a exploité le potentiel des donnée non structuré, cette génération a été largement étudiée et utiliser dans le web actuel.

L'évolution du web 2.0 vers le web sémantique a présenté un besoin pour le développement d'une nouvelle génération qui est les systèmes question réponse pour les données liée (SQRDL). Les SQRDL utilise les donnée liée notamment les ontologies pour réponde aux questions des utilisateurs.

On a présenté aussi une évaluation pour mesurer les performances des systèmes question réponse.

Chapitre 2

Classification des questions

1) Introduction

On regroupe sous le vocable de traitement automatique du langage naturel (TALN) l'ensemble des recherches et développements visant à modéliser et reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication. Ce chapitre donnera un aperçu sur le traitement automatique du langage naturel.Nous allons présenter dans un premier temps la définition de TALN puis un bref historique. Nous aborderons ensuite l'objectif du TALN en précisant l'analyse et génération. Les différents niveaux d'analyse des textes seront exposés. Une brève des niveaux de traitement. Nous finirons le chapitre par citer les différents outils du TALN.

2) NLP (TALN)

Le **NLP, Natural Language Processing** ou **Traitement Automatique du Langage en français**, désigne l'ensemble des tâches permettant à un ordinateur de traiter des données en langage humain. Il s'agit donc d'une discipline informatique à part entière qui recouvre de nombreux sujets et méthodes, qui sont à l'origine notamment des moteurs de recherche. Certains auteurs distinguent des tâches dites de "bas-niveau" permettant une représentation informatique du texte par un ordinateur, des tâches dites de "haut-niveau" permettant à la machine de "comprendre" le texte.

Un traitement automatique du langage naturel (TALN) est l'ensemble des méthodes et des programmes qui permettent un traitement par l'ordinateur des données langagières, il est aussi une suite d'actions ou calculs à faire effectuer par la machine. Il a pour objectif de traiter des données linguistiques (textes) exprimées dans une langue dite "naturelle". En plus la conception de programmes capables de traiter automatiquement des données linguistiques de type textes écrits : dialogues écrits ou oraux et des unîtes linguistiques (mots, phrases, énonces, ...). Aussi TALN est une discipline s'appliquant au domaine de l'informatique et du langage. Il est utilisé par exemple pour les traductions, la reconnaissance vocale ou encore les réponses automatiques aux questions. Ces domaines représentent des défis majeurs, car les mots du langage sont souvent traités un à un par l'ordinateur.

Grâce au traitement du langage naturel, une cohérence tente d'être apportée aux textes en s'attachant au sens des phrases et formules. Ces avancées ne sont pas uniquement utiles pour les traducteurs ou chat bots mais aussi lorsque les ordinateurs exécutent des

ordres oraux ou communiquent de manière vocale afin de faciliter par exemple la communication pour les personnes aveugles. Pour pouvoir résumer des textes longs, ou extraire des informations précises, les ordinateurs ont besoin également de comprendre la cohérence linguistique des textes.

En Etats-Unis, 1949, Warren Weaver dans Memorandum parle de computer translation et de mechanization of the translation problem, aborde les problèmes de l'automatisation du traitement du langage, propose de résoudre les ambiguïtés syntaxiques et sémantiques en utilisant la redondance du langage écrit dans le cadre de la théorie de l'information. Dans les années 1960, apparition du terme Automatic [20].

2.1- Historique

Dans les Computational Linguistics, deux courants naissent : les recherches purement pragmatiques avec acceptation d'une marge d'erreur devant aboutir à des résultats concrets dans des délais limités, mais dont l'exploitation effective est essentiellement liée à l'introduction rapide et peu onéreuse des données en machine ; des travaux relevant de la recherche fondamentale, soit à tendance linguistique, soit à tendance mathématisante. 1967, conférence internationale du traitement automatique des langues, organisée par le CETA, à Grenoble : analyse automatique des langues naturelles, analyse statistique et sémantique des données linguistiques, théorie algébrique des langages. Dans les années 1970, une rencontre entre automatisation du langage et intelligence artificielle fait naitre le terme Natural Language Processing (NLP). Ce dernier s'installe dans les années 1980 et semble bien établi dans les années 1990.

2.2- Objectif du TALN

L'objectif du TALN est la conception de logiciels, capables de traiter de façon automatique des données linguistiques, c'est-à-dire des données exprimées dans une langue (dite "naturelle"). Ces données linguistiques peuvent être des textes écrits, ou bien des dialogues écrits ou oraux, ou encore des unités linguistiques de taille inférieure à ce que l'on appelle habituellement des textes (par exemple : des phrases, des énoncés, des groupes de mots ou simplement des mots isolés). En fonction des objectifs attendus d'un système de TALN, on peut le classer dans l'un des domaines suivants :

Analyse et génération.

- **L'analyse** : Le processus d'analyse consiste à démarrer de la structure de surface (locution ou texte écrit) pour arriver à la structure profonde équivalente. Plus précisément, l'ordinateur doit pouvoir interpréter un texte écrit dans une langue naturelle afin d'en

obtenir une représentation formelle. Pour ce faire, il doit pouvoir analyser le langage à différents niveaux : morphologique, syntaxique, sémantique, pragmatique, l'ordinateur doit donc faire appel à des connaissances linguistiques importantes (grammaires, dictionnaires, etc.) ainsi qu'à des connaissances pragmatiques (contexte immédiat, expérience, connaissances du monde, etc.).

- **La génération** : Le processus de génération est le processus inverse : il consiste à passer de la structure profonde à la structure de surface. Plus précisément, l'ordinateur doit être capable de passer de la représentation formelle d'un énoncé à une formulation en langue naturelle. Ceci implique la même connaissance que pour l'analyse, mais avec des problèmes spécifiques à l'énonciation.

2.3. Les niveaux d'analyse d'un texte

On distingue plusieurs niveaux d'analyse d'un texte:

Le niveau morphologique ou lexical qui s'intéresse à la décomposition d'un texte en différents mots (ex : en Français découpage selon les espaces).

Le niveau syntaxique qui définit comment les mots est agencé dans une phrase, et l'analyse de la forme d'un texte.

Le niveau sémantique qui correspond à la signification des mots et des groupes de mots (analyse du sens d'un texte (ex : typage)) [23].

Le niveau pragmatique grâce auquel, le contexte est pris en compte. (Voir la figure 5)

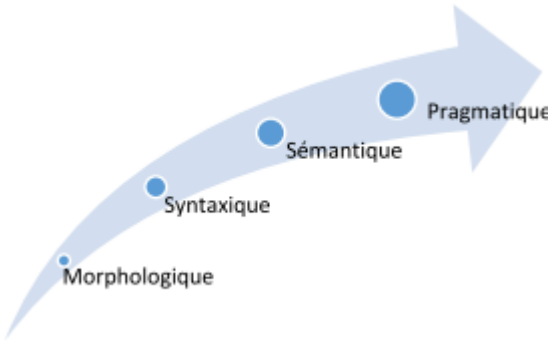

Figure 5:Les niveaux d'analyse d'un texte.

2.3.1. Morphologie

La morphologie est l'étude de la forme des mots (de leur flexion – indications de cas, genre, nombre, mode, temps, etc. de leur d´éviration préfixes, suffixes, infixes et de leur composition mots composés). Sous l'appellation de morphosyntaxe, elle représente

également l'étude des règles de combinaison des morphèmes (unités minimales de sens) selon la configuration syntaxique de l'énoncé. En pratique, dans le cadre du traitement automatique de la langue, l'analyse morphologique consiste à segmenter le texte en unités élémentaires (tokenisation) et à déterminer les différentes caractéristiques de ces unités.

2.3.2. Analyseur syntaxique

L'analyse syntaxique est l'étude de la structure de la phrase, dans le but de définir comment les lexèmes sont organisés et quelles fonctions ont les mots qui servent à les mettre en relation. Les lexèmes sont réunis en syntagmes, des groupes de mots dont la fonction est connue. Il par constituants divise tout d'abord la phrase en plusieurs groupes de mots appelés syntagmes. Un syntagme est un intermédiaire entre l'ensemble global qu'est la phrase et la division unitaire que sont les mots. Il permet de représenter, sous forme symbolique ou graphique, la ou les structures syntaxiques d'un texte. En d'autres termes, il s'agit de la mise en évidence des structures d'agencement des catégories grammaticales (nom, verbe, adjectif, etc.), afin d'en découvrir les relations formelles ou fonctionnelles (par exemple, sujet, verbe et complément).

2.3.2.1. Notion de syntagme

Un syntagme est un ensemble de mots formant une seule unité catégorielle et fonctionnelle, mais dont chaque constituant, parce que dissociable (contrairement au mot composé), conserve sa signification et sa syntaxe propres. Un syntagme constitue donc une association occasionnelle, libre, alors que le mot composé est une association permanente (lorsqu'un syntagme se fige, il devient bien sûr un composé détaché, soit une locution). Il s'agit donc d'un groupe de mots formant une unité à l'intérieur de la phrase. Dans le cadre de l'analyse syntaxique d'une phrase, il s'agit d'une segmentation en unités fonctionnelles appelées syntagmes. Par exemple, on peut citer les types de syntagme suivants : syntagme nominal, syntagme verbal, syntagme adjectival, etc. Les arbres de dérivation/syntaxique sont un moyen de représentation des syntagmes. Exemple 1 : pour la phrase « Le facteur apporte une lettre », l'arbre de dérivation est donné dans la figure 2.3.

Soit la grammaire qui donne d'abord le sens des abréviations

(SV, SN, S, V, Det ….)

P→ SN | SV SN → Det | N SV → V| SN pp PP → DT | N

2.3.3 - La sémantique

La sémantique vise à l'étude de sens hors contexte. Le traitement sémantique prend comme unité d'analyse de la phrase, et conduit à représenter sa partie significative. L'analyseur sémantique d'écrit le sens des mots de la phrase ; ces mots sont identifiées par l'analyse morphologique, et regroupes en structures par l'analyse syntaxique.

2.3.4 - L'analyse pragmatique

L'analyse pragmatique permet d'utiliser les connaissances pragmatiques afin d'interpréter des situations du monde réel. Cette étape est importante pour le processus de compréhension d'un texte ou d'une phrase, elle représente le lien entre l'analyse linguistique et le monde réel.

2.4- Application du TALN Traduction automatique(TA) :

Il est probable que la TA fasse l'objet d'améliorations importantes dans les années à venir.

Correction orthographique : Intégrée à toute application informatique impliquant la rédaction, correction basée sur des lexiques. Exemple: traitement de texte, courrier électronique, navigateur Internet (zone de saisie) (figure 6).

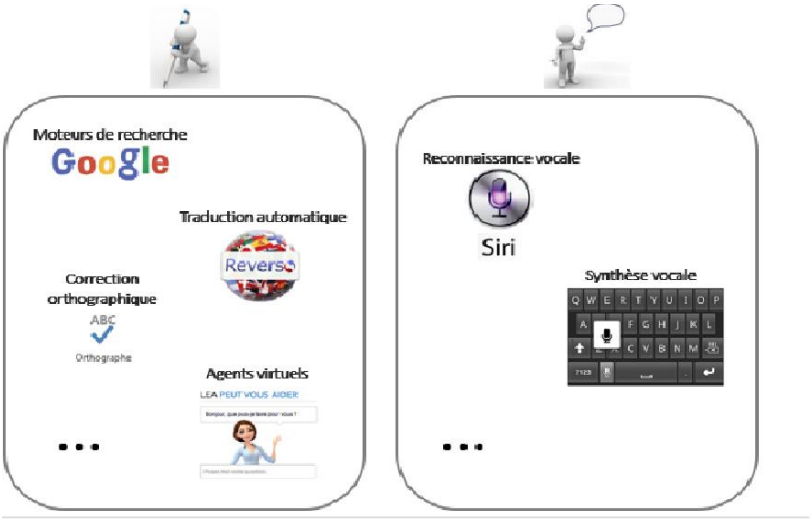

Figure 6:les applications du TALN

Correction grammaticale : Les meilleures applications fonctionnent bien mais sont payantes, actuellement aucune application libre pour le français.

Atelier d'aide à la rédaction : Orthographe, grammaire, style…, Exemple : Antidote pour le français.

La reconnaissance de la parole :Discipline ayant fait des progrès considérables , Grandes étapes : {Segmentation du flux continu de paroles en unités discrètes identification du phonème correspondant à chaque unité ,regroupement des unités pour constituer des mots, prise en compte de la syntaxe pour finaliser le texte écrit } , logiciels de dictée vocale (Via Voice, Dragon Dictate…) , reconnaissance de la parole ou commande vocale (Reconnaissance vocale de Windows, Systèmes de navigation routière GPS, Smartphone…) , prototype Google de sous-titrage automatique de Youtube.

Synthèse de la parole : Créer de la parole artificielle à partir d'un texte quelconque. Ces systèmes ont largement franchi le seuil de l'intelligibilité permettant leur utilisation, difficultés : désambiguïsation des homographes hétérophonies, gestion de la prosodie (intonation, rythme et intensité)….

2.5 - Les outils de TALN

Nous citons les outils suivants :

2.5.1. Le Natural Langage Toolkit (NLTK)

Est un ensemble d'outils TALN en langage Python. L'outil propose un accès à plus de 100 corpus de textes, parmi lesquels des textes en anglais, portugais, polonais, néerlandais, catalan et basque. De plus, le kit peut effectuer le traitement de différents textes, comme l'étiquetage morphosyntaxique, l'arbre syntaxique, la segmentation (tokenisation en anglais, ce qui constitue souvent la première étape du TALN) et la synthèse de texte.

Le kit d'outils TALN comporte également une introduction à la programmation et une documentation détaillée. Il est ainsi bien adapté aux étudiants, doctorants et chercheurs.

2.5.2. Stanford NLP Group Software (StanfordCoreNLP)

Est l'un des groupes de recherche les plus importants dans le domaine du traitement automatique du langage naturel. De nombreux outils sont proposés. Ils permettent de définir la forme de base des mots (segmentation en unité), la fonction des mots (étiquetage morphosyntaxique) et la structure des phrases (arbre syntaxique). De plus, il existe des outils pour les processus compliqués comme le deeplearning pour lequel le contexte de la phrase est pris en compte. L'outil StanfordCoreNLP présente la plupart des fonctions de base. L'ensemble des programmes du Standford NLP sont écrits en langage Java et sont disponibles en français, anglais, allemand, espagnol et chinois.

2.5.3.CSLU

Pour le traitement du langage vocal, on trouve le kit CSLU (Center for SpokenLanguageUnderstanding). Ces outils présentent, entre autres fonctions, la reconnaissance vocale et la retransmission de textes à l'oral (par voix de synthèse). Ils comprennent également des outils d'entraînement avec lesquels les enfants peuvent apprendre des nouveaux mots de vocabulaire, et avec lesquels les personnes sourdes peuvent s'exercer à parler. Les outils sont ainsi adaptés aux jeunes élèves, aux étudiants, aux chercheurs et bien sûr à toute autre personne intéressée.

2.5.4. Visualtext

C'est un ensemble d'outils écrits dans un langage de programmation propre au TALN :

Le langage NLP++. Ce langage de programmation a surtout été développé pour ce que l'on appelle les analystes DeepText. Visualtext sert à extraire des informations depuis une grande quantité de textes. Il permet par exemple de résumer des textes longs mais aussi de regrouper des évènements sur un thème précis à partir de plusieurs sites Web et de créer un aperçu. Visualtext peut être utilisé gratuitement pour des fins non-commerciales.

3) Machine Learning

3.1 Introduction

L'apprentissage automatique (ou artificiel) (machine-learning en anglais) est un des champs d'étude de l'intelligence artificielle. Commençons par la définition de l"AAAI1 et celle fournie dans l'avant-propos de (Cornuéjols et al. 2002). L'apprentissage artificiel fait référence à la capacité d'un système à acquérir et intégrer de façon autonome des connaissances. Cette notion englobe toute méthode permettant de construire un modèle de la réalité à partir de données, soit en améliorant un modèle partiel ou moins général, L'apprentissage automatique fait référence au développement, l'analyse et l'implémentation de méthodes qui permettent à une machine (au sens large) d'évoluer et de remplir des tâches associées à une intelligence artificielle grâce à un processus d'apprentissage. Cet apprentissage permet d'avoir un système qui s'optimise en fonction de l'environnement, les expériences et les résultats observés. Voyons quelques exemples. La capacité d'apprentissage est une caractéristique des êtres vivants. De la naissance à l'âge adulte, les êtres vivants acquièrent de nombreuses capacités qui leur permettent de

survivre dans leur environnement. L'apprentissage d'un langage, de l'écriture et de la lecture sont de bons exemples des capacités humaines, et des phénomènes mis en jeu : apprentissage par cœur, apprentissage supervisé par d'autres êtres humains, apprentissage par généralisation. Une application classique de l'apprentissage artificiel est la reconnaissance de caractères manuscrits, tels qu'ils apparaissent sur une enveloppe. La difficulté tient au fait que la variété des formes rencontrée est infinie. L'apprentissage par cœur n'est pas possible, et il faut donc être capable de généraliser à partir d'un ensemble d'exemples de caractères (figure 2.3).

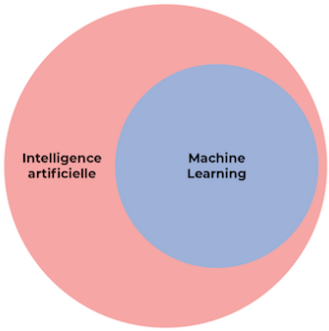

Figure 7:Machine Learning est une sous-discipline de l'IA.

- Quelques mots d'Histoire...

L'apprentissage artificiel est une discipline jeune, à l'instar de l'Informatique et de l'Intelligence artificielle. Il se situe au carrefour d'autres disciplines: philosophie, psychologie, biologie, logique, mathématique .les premières études remontent à des travaux de statistique dans les années 1920. C'est après la seconde guerre mondiale que les premières expériences deviennent possibles. Se développent ensuite dans les années 1960 les approches connexionnistes avec des perceptrons, et la reconnaissance des formes. La mise en évidence des limites du perceptron simple arrête toutes les recherches dans ce domaine jusqu'à la renaissance dans les années 1980. Les années 1970 sont dominées par des systèmes mettant l'accent sur les connaissances, les systèmes experts, Les limites de tels systèmes se font sentir dans les années 1980, pendant lesquelles a lieu le retour du connexionnisme avec un nouvel algorithme d'apprentissage. Les mathématiciens commencèrent à s'éloigner du cadre cognitif de l'apprentissage pour envisager le problème sous l'angle de l'optimisation, pendant qu'apparaissaient de nouvelles méthodes comme les arbres de décision ou l'induction de programmes logiques. L'influence de la théorie statistique de l'apprentissage s'est réellement fait sentir dans les années 1990, Dans le présent chapitre, nous commencerons par introduire les concepts fondamentaux du

Machine Learning supervisé dans le cas de la classification. Nous validerons ensuite cette approche en effectuant une analyse statistique de l'apprentissage, principalement basée sur la théorie de Vapnik. Finalement, nous présenterons les principales étapes de la mise en pratique de la méthodologie ML. Dans la dernière section de ce chapitre nous présentons les différentes méthodes de classification.

3.2 Concepts et Sources de l'apprentissage automatique :

L'apprentissage de l'être humain se compose de plusieurs processus qu'il est difficile précisément à décrire. Les facultés d'apprentissage chez l'humain lui ont conféré un avantage évolutif déterminant pour son développement. Par " faculté d'apprendre " on entend un ensemble d'aptitudes comme :

- L'obtention de la capacité de parler en observant les autres.
- L'obtention de la capacité de lire, d'écrire, d'effectuer des opérations arithmétiques et logiques avec l'aide d'un tuteur.
- L'obtention d'habilités motrices et sportives en s'exerçant.

3.2.1 Qu'est-ce que l'apprentissage automatique

La faculté d'apprendre de ses expériences passées et de s'adapter est une caractéristique essentielle des formes de vies supérieures. Elle est essentielle à l'être humain dans les premières étapes de la vie pour apprendre des choses aussi fondamentales que reconnaître une voix, un visage familier, apprendre à comprendre ce qui est dit, à marcher et à parler. L'apprentissage automatique est une tentative de comprendre et reproduire cette faculté d'apprentissage dans des systèmes artificiels. Il s'agit, très schématiquement, de concevoir des algorithmes capables, à partir d'un nombre important d'exemples (les données correspondant à "l'expérience passée"), d'en assimiler la nature afin de pouvoir appliquer ce qu'ils ont ainsi appris aux cas futurs.

3.2.2 Définition

Un programme d'ordinateur est capable d'apprendre à partir d'une expérience E et par rapport à un ensemble T de tâches et selon une mesure de performance P, si sa performance à effectuer une tâche de T, mesurée par P, s'améliore avec l'expérience E.

3.2.3 Modélisation

L'apprentissage automatique d'une machine toujours concerne un ensemble de tâches concrètes - T. Pour déterminer la performance de la machine, on utilise une mesure

de la performance P. La machine peut avoir à l'avance un ensemble d'expérience E ou elle va enrichir cet ensemble plus tard (figure 8).

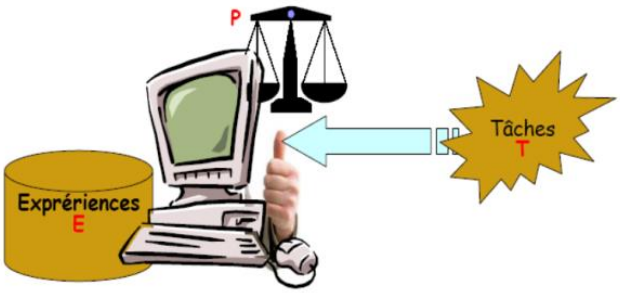

Figure 8:schéma de modélisation d'une machine d'apprentissage.

Donc, l'apprentissage automatique pour la machine est qu'avec l'ensemble de tâches T que la machine doit réaliser, elle utilise l'ensemble d'expériences E telle que sa performance sur T est améliorée.

3.2.4 Domaines d'applications de l'apprentissage automatique:

L'apprentissage automatique s'applique à un grand nombre d'activités humaines et convient en particulier au problème de la prise de décision automatisée (figure 9). Il s'agira, par exemple:

¬ D'établir un diagnostic médical à partir de la description clinique d'un patient;

¬ De donner une réponse à la demande de prêt bancaire de la part d'un client sur la base de sa situation personnelle;

¬ De déclencher un processus d'alerte en fonction de signaux reçus par des capteurs ;

¬ De la reconnaissance des formes;

¬ De la reconnaissance de la parole et du texte écrit;

¬ De contrôler un processus et de diagnostiquer des pannes;

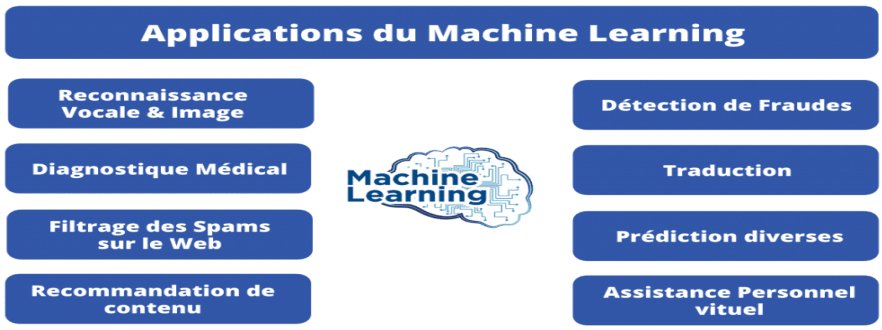

Figure 9:Applications de l'apprentissage automatique.

3.2.5 Mémoriser n’est pas généraliser :

Le terme apprentissage dans la langue courante est ambigu. Il désigne aussi bien l'apprentissage “par cœur” d’une poésie, que l’apprentissage d’une tâche complexe telle que la lecture. Clarifions la distinction :

– Le premier type d’apprentissages correspond à une simple mémorisation. Or les ordinateurs contemporains, avec leurs mémoires de masse colossales, n’ont aucune difficulté à mémoriser une encyclopédie entière, sons et images inclus.

– Le second type d’apprentissages se distingue fondamentalement du premier en cela qu’il fait largement appel à notre faculté de généraliser. Ainsi pour apprendre à lire, on doit être capable d’identifier un mot écrit d’une manière que l’on n’a encore jamais vue auparavant.

3.3 Types d'apprentissage

Les algorithmes d'apprentissage peuvent se catégoriser selon le mode d'apprentissage qu'ils emploient (figure 10) :

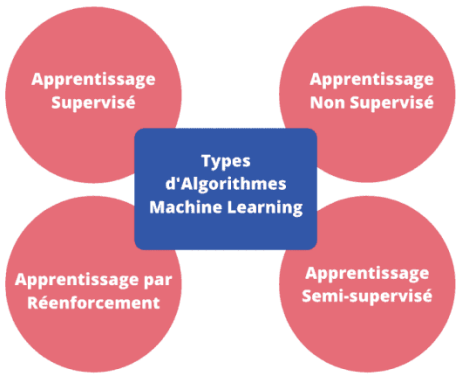

Figure 10:types-d’algorithmes-d’apprentissage-automatique.

3.3.1 L'apprentissage supervisé

Si les classes sont prédéterminées et les exemples connus, le système apprend à classer selon un modèle de classement ; on parle alors d'apprentissage supervisé (ou d'analyse discriminante). Un expert (ou oracle) doit préalablement correctement étiqueter des exemples. L'apprenant peut alors trouver ou approximer la fonction qui permet d'affecter la bonne « étiquette » à ces exemples. Parfois il est préférable d'associer une donnée non pas à une classe unique, mais une probabilité d'appartenance à chacune des classes prédéterminées (on parle alors d'apprentissage supervisé probabiliste). L'analyse discriminante linéaire ou les SVM sont des exemples typiques. Autre exemple : en fonction de points communs détectés avec les symptômes d'autres patients connus (les « exemples

»), le système peut catégoriser de nouveaux patients au vu de leurs analyses médicales en risque estimé (probabilité) de développer telle ou telle maladie.

3.3.2 L'apprentissage non-supervisé

Quand le système ou l'opérateur ne dispose que d'exemples, mais non d'étiquettes, et que le nombre de classes et leur nature n'ont pas été prédéterminés, on parle d'apprentissage non supervisé (ou clustering). Aucun expert n'est disponible ni requis. L'algorithme doit découvrir par lui-même la structure plus ou moins cachée des données. Le système doit ici dans l'espace de description (la somme des données) cibler les données selon leurs attributs disponibles, pour les classer en groupe homogènes d'exemples. La similarité est généralement calculée selon la fonction de distance entre paires d'exemples. C'est ensuite à l'opérateur d'associer ou déduire du sens pour chaque groupe. Divers outils mathématiques et logiciels peuvent l'aider. On parle aussi d'analyse des données en régression. Si l'approche est probabiliste (c'est à dire que chaque exemple au lieu d'être classé dans une seule classe est associé aux probabilités d'appartenir à chacune des classes), on parle alors de « soft clustering » (par opposition au « hard clustering ») . Exemple : Un épidémiologiste pourrait par exemple dans un ensemble assez large de victimes de cancers du foie tenter de faire émerger des hypothèses explicatives, l'ordinateur pourrait différencier différents groupes, qu'on pourrait ensuite associer par exemple à leur provenance géographique, génétique, à l'alcoolisme ou à l'exposition à un métal lourd ou à une toxine telle que l'aflatoxine.

3.3.3 L'apprentissage semi-supervisé

Effectué de manière probabiliste ou non, il vise à faire apparaitre la distribution sous-jacente des « exemples » dans leur espace de description. Il est mis en œuvre quand des données (ou « étiquettes ») manquent... Le modèle doit utiliser des exemples non-étiquetés pouvant néanmoins renseigner. Exemple : En médecine, il peut constituer une aide au diagnostic ou au choix des moyens les moins onéreux de tests de diagnostics.

3.3.4 L'apprentissage partiellement supervisé (probabiliste ou non)

Quand l'étiquetage des données est partiel. C'est le cas quand un modèle énonce qu'une donnée n'appartient pas à une classe A, mais peut-être à une classe B ou C (A, B et C étant 3 maladies par exemple évoquées dans le cadre d'un diagnostic différentiel).

3.3.5 L'apprentissage par renforcement

L’algorithme apprend un comportement étant donné une observation. L'action de l'algorithme sur l'environnement produit une valeur de retour qui guide l'algorithme.

3.4 Les algorithmes utilisés

- Les machines à vecteurs support
- Le boosting
- Les réseaux de neurones pour un apprentissage supervisé ou non-supervisé
- La méthode des k plus proches voisins pour un apprentissage supervisé
- Les arbres de décision
- Les méthodes statistiques comme par exemple le modèle de mixture gaussienne
- La régression logistique
- L'analyse discriminante linéaire
- La logique floue
- Les algorithmes génétiques et la programmation génétique.

4. Conclusion

Dans ce chapitre, nous avons présenté la notion du TALN, ses différentes applications et techniques en mettant l'accent sur les niveaux d'analyses ce qui apporte au texte un premier niveau de compréhension à partir d'une analyse syntaxique et grammaticale. Le résultat de ces outils de traitement du langage naturel servira ensuite d'entrée pour des fonctionnalités plus poussées comme l'analyse sémantique et la découverte d'informations.

Dans notre travail, nous utilisons ces outils afin de simplifier les phrases (réponses de l'apprenant et réponse type) en phrases plus simples et plus réduites. Les règles de réduction proposées s'appuient sur la segmentation du texte en phrases et chaque phrase en tokens, chacun possède son POS (Part Of Speech). Ensuite, en se basant sur ces POS, on simplifie, réduit et transforme chaque phrase du texte en un triplet SVO : (Sujet, Verbe, Objet) supposés porteurs de l'information des phrases dont ils sont extraits. Après la phase de simplification des phrases (celles de la réponse de l'apprenant ainsi que celles proposées comme corrigé type), les SVO générés contribuent dans la phase de calcul de similarité sémantique entre la réponse de l'apprenant et le corrigé type fourni. Le calcul repose sur les méthodes proposées par l'ontologie WordNet [18].

Chapitre 3

AQALD: Systeme question response

Arable pour les données liée

1. Introduction

La réponse automatique aux questions est devenue une direction de recherche en traitement automatique du langage naturel. Son but est de chercher une réponse précise et concise à une forme libre question factuelle à partir d'une grande collection de données textuelles, plutôt qu'un document complet, jugé pertinent comme dans une information standard tâches de récupération. Bien que différents types de questions les répondeurs ont des architectures différentes, la plupart elles suivent un cadre dans lequel la classification des questions joue un rôle important. De plus, certaines études ont a démontré que la performance de la classification des questions a une influence significative sur la performance globale d'un Système questions et réponse. La tâche de classification des questions est de prédire le type d'entité de la réponse d'une naturelle question de langue6. Par exemple, pour la question « Où est la Tour Eiffel ? », la tâche de la classification des questions est d'étiquette de retour « emplacement », donc la réponse à cette question est une entité nommée de type « emplacement ». Puisque nous prédisons le type de la réponse, la classification des questions est également appelée réponse prédiction de type.

De nombreuses études ont abordé ce problème, elles appartiennent à l'approche basée sur des règles ou basée sur l'apprentissage automatique approche. Dans ce mémoire, nous suivons la machine approche d'apprentissage et prêter attention à l'importance

D'extraction et de sélection de caractéristiques. Du point de vue d'apprentissage automatique, nous pouvons facilement formuler cette tâche comme un problème de classification. Il y a plusieurs supervisés méthodes d'apprentissage utilisées telles que les voisins les plus proches (NN), Naive Bayes (NB), Arbre de décision (DT), Réseau clairsemé de Winnows (SNoW) et Support Vector Machines (SVM). Cependant, comme exprimé à partir des résultats expérimentaux dans les précédentes études, les ensembles de fonctionnalités affectent beaucoup la qualité de la question classification.

Selon des études antérieures, divers types de caractéristiques ont fait l'objet d'une enquête. Les types les plus courants sont sac de mots et de n-grammes qui ont été utilisés dans toutes les études. D'autres études ont tenté d'enrichir l'ensemble de fonctionnalités en ajoutant plus d'informations linguistiques dans le discours des balises ou des mots clés, ou même des caractéristiques sémantiques. Cependant, de notre observation, combiner toutes les caractéristiques n'est pas toujours la meilleure solution pour toutes les questions. Par conséquent, dans ce chapitre nous allons donner une enquête expérimentale pour trouver les meilleurs ensembles de fonctionnalités correspondant aux différents groupes de

questions. De plus, nous extrayons également un nouveau type de fonctionnalités basées sur des modèles de questions. Ces nouvelles fonctionnalités sont ensuite intégrées aux ensembles de fonctionnalités existants et reçoivent meilleurs résultats de classification. Nous avons testé notre proposition ensembles de fonctionnalités utilisant un classificateur SVM qui est expérimental montré pour obtenir les meilleurs résultats.

Le reste de ce chapitre est organisé comme suit :

Architecture des systèmes de réponse aux questions, Puis donne une approche pour classer une question, et la Comparaison des algorithmes de classification et Techniques PNL pour l'analyse des questions et classification et la dernière parte

L'évaluation est perspective de notre système

2. Arrière-plan

Système de réponse aux questions

« Système de questions-réponses » ou « Système de réponse aux questions », abrégé en SQR, est un domaine de recherche en pleine croissance qui rassemble la recherche d'informations (RI), l'information

Système (SI), Web sémantique, Intelligence Artificielle (IA), Machine Learning (ML) et le traitement du langage naturel (NLP). Techniques et méthodes développés pour SQR sont, à leur tour, utilisé dans de nombreux domaines interdépendants, tels que Document Récupération, reconnaissance d'entités nommées (NER), génération d'ontologies, chatbot Agent de conversation, etc.

La construction de systèmes de questions-réponses a commencé avec un système appelé BASEBALL. En 1965, présente quinze des systèmes de questions-réponses mis en place pour répondre automatiquement aux questions en anglais. Ces premiers systèmes mis en œuvre se sont concentrés sur des domaines et des problèmes spécifiques[14].

Plus tard, le domaine a évolué vers de nouvelles tendances, grâce à la disponibilité de L'informations, série de conférences d'évaluation organisées et tâches de la question systèmes de réponse. Les entreprises les plus importantes, telles que Google, Yahoo, Microsoft et IBM, se sont intéressées à la mise en œuvre de projets, qui témoignent de la popularité croissante de ce secteur.

La première génération de systèmes de réponse aux questions était Natural Language Interface aux bases de données (NLIDB). Ce type de système interroge les bases de données en utilisant une requête en langage naturel au lieu du langage de requête formel, tel que SQL.NLIDB permet au grand public d'explorer la base de données sans avoir à

faire à un langage de requête formel. La recherche NLIDB conduit à une architecture basée sur deux composants : le composant linguistique et le composant base de données (Figure 11).

Avec l'évolution de l'informatique et des technologies Web, de plus en plus déstructurées les données étaient disponibles sur le Web. Le système textuel de réponse aux questions vise à générer une réponse à une requête en langage naturel à partir de sources textuelles, telles que des documents et pages Web HTML. Ce type de SQR suit une architecture générale

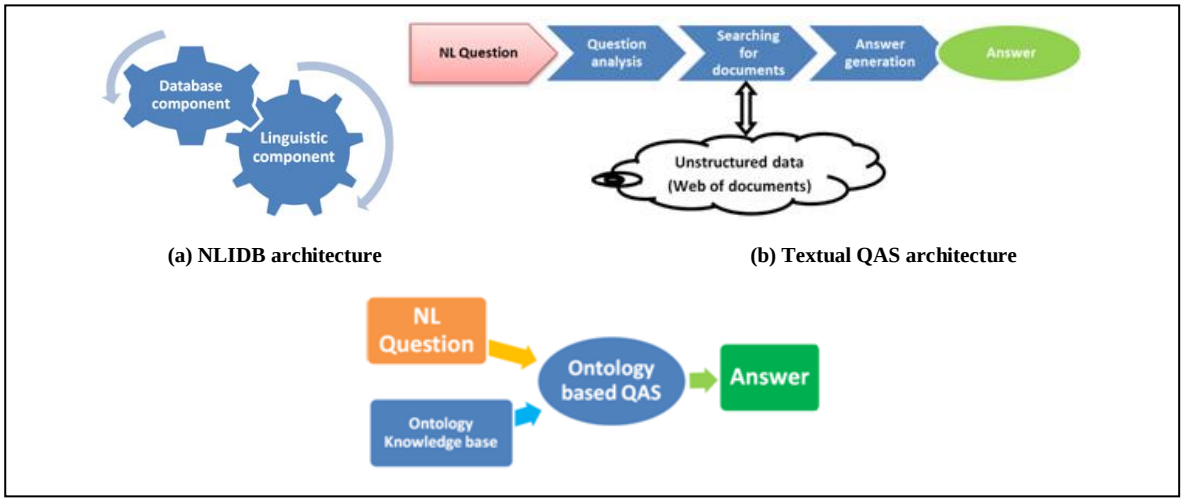

Figure 11:Architectures des systèmes de réponse aux questions.

(Figure 11) basé sur trois modules : analyse des questions, recherche de documents et génération de réponses.

L'évolution des technologies du Web sémantique a conduit à étendre le Web des documents avec les langages de données liées et d'ontologie, tels que OWL, RDF, RDFS et SPARQL. Mutation de SQR vers une nouvelle génération nommée Question Answering over ,Les données liées (QALD) doivent traiter des données structurées et valides. La plupart des QALD utilise des correspondances entre les termes de la question et la base de connaissances de l'ontologie (KB) entités et propriétés pour construire une requête SPARQL, selon le type de question[21]. L'architecture de QALD est illustrée à la (figure 11).

3. Travaux connexes

A Il existe trois approches pour classer les questions : approche basée sur des règles, approche d'apprentissage automatique et approche hybride, qui combine une approche basée sur des règles et une approche basée sur l'apprentissage. La première est définie par des règles spécifiques en fonction des modèles. La seconde approche, l'apprentissage automatique, classe les questions après une étape d'apprentissage qui nécessitent un ensemble de données annotées.

Dans notre mémoire, nous avons adopté l'approche d'apprentissage automatique, car elle :

- Est le plus utilisé
- Couvrir tous les types de questions
- Flexible pour les nouvelles données
- Moins compliqué que l'approche basée sur des règles

De nombreux algorithmes ont été appliqués à la classification de texte. La plupart des études ont été consacrées à l'anglais et à d'autres langues latines. Cependant, très peu de recherches ont été menées sur le texte arabe :

- El Koudri [16] a classé automatiquement les documents Web arabes par Naive Bayes (NB) qui est un algorithme d'apprentissage automatique. Des expériences de validation croisée ont été utilisées pour évaluer les résultats du NB. La précision de la catégorisation varie d'une catégorie à l'autre avec une précision moyenne toutes catégories confondues de 68,78 %.
- Entropie maximale (ME) utilisée pour classer les articles de presse arabes. La précision de la classification était de 80,41% et 62,7% par Sawaf sans aucune analyse morphologique [12].
- Meshleh met en œuvre un système de classification de texte basé sur une machine à vecteurs de support (SVM) pour les articles en langue arabe. Il a utilisé un corpus d'archives de journaux arabes en ligne, notamment Al-Jazeera, Al-Nahar, Al-Hayat, Al-Ahram et Al-Dostor. Le système montre une classification élevée efficace pour l'ensemble de données arabe en terme de F-mesure (F=88,11).
- Harrag présente les résultats de la classification de documents texte arabes en utilisant un algorithme d'arbre de décision (DT). L'étude a conclu que l'efficacité du classificateur amélioré est très bonne et donne une précision de généralisation de l'ordre de 0,93 pour le corpus scientifique et de 0,91 pour le corpus littéraire.

• Réseau de neurones artificiels (ANN) pour la classification est également utilisé, pour classer les documents arabes. Pour le corpus utilisé, la performance atteint 88,75%.

• Halees[15] dans a réalisé une étude comparative pour six classificateurs : ME, NB, DT, ANN, SVM et KNN, avec le même ensemble de données. Il a trouvé que les performances de Naïve Bayes sont les meilleures (F1 = 91,81), les performances de l'entropie maximale, de la machine à vecteurs de support et de l'arbre de décision sont acceptables.

Table 1:Comparaison des algorithmes de classification

Reference	Used classifier	Accuracy or F-measure
El Koudri [1]	Naive Bayes	68.78 %
El-Halees [2]	Maximum entropy	80.41%
Sawaf [3]	Maximum entropy	62.7%
Meshleh [4]	Support Vector Machines	F=88.11
Harrag [5]	Decision tree	93%
Harrag [5]	Artificial Neural Network	88.33%
Halees [6]	Naïve Bayes	F=91.81

Dans tous les systèmes précédents, chaque auteur utilise son propre ensemble de données, pour cette raison, nous ne pouvons pas prendre de décision sur le meilleur classificateur de questions en arabe.

Concernant le non disponibilité des ressources arabes, chaque auteur utilise son propre jeu de données pour tester sa méthode. Pour cette raison, nous ne pouvons pas faire une étude comparative pour les SQR arabes existants. On propose de construire un jeu de données, accessible à tous, qui couvre tous les types de questions dans toutes les catégories. A ce stade, nous pouvons mesurer l'évolution des QAS arabes [19].

4. Analyse des questions

L'analyse des questions est une composante primordiale dans la plupart des SQR. Ce composant vise à extraire des informations significatives de la question, telles que la question type (classe de question), focus, entités nommées et relation. Dans la littérature, les chercheurs proposent plusieurs approches pour traiter la question d'entrée. Celles-ci Les approches peuvent être classées en deux catégories principales : l'approche TLN basée sur des règles et approche ML. L'approche TLN basée sur des règles utilise différents niveaux

de techniques TLN pour l'analyse de la question, telles que l'analyse morphologique, l'analyse syntaxique, analyse sémantique, pragmatique et analyse du discours. En raison de la disponibilité du grand jeu de données extrait du Web, les approches ML sont aujourd'hui les plus Techniques prometteuses dans différentes tâches de TLN, y compris la tâche de réponse aux questions. La méthode hybride utilise une combinaison de techniques NLP et ML pour traiter le question dans la tâche SQR. Notre système proposé utilise une approche TLN basée sur des règles pour analyser la question et extraire les informations suivantes : question classe, mots-clés et ressource. Dans ce qui suit, nous décrivons les différentes étapes de la composante d'analyse des questions.

5. Prétraitement

La première étape consiste à prétraiter la question par tokenisation : pour scinder le texte dans les unités élémentaires et normalisation : pour éviter l'orthographe commune erreurs dans la question d'entrée parce que la langue arabe peut être écrite en différentes manières. La correction des fautes d'orthographe les plus courantes passe par la normalisation de l'arabe Alif "ا" et Ya "ي" (Habash, 2010). Ensuite, les mots de la question d'entrée sont étiquetés par le marquage POS afin d'identifier la catégorie grammaticale de chaque terme (étiquettes POS). Le jeu de balises utilisé dans notre étude peut-être trouvé dans [14].

6. Classification des questions

Une des informations essentielles extraites de la question est sa classe (type, Catégorie). La classification des questions consiste à étiqueter la question d'entrée selon à la taxonomie des questions spécifiques. La classification des questions explore les points suivants informations[11]:

1. La classe de question peut prédire une contrainte sémantique sur la réponse souhaitée.

2. La classe de question détermine la structure syntaxique de la question.

3. La classe de questions peut aider au développement du modèle de réponse.

Les recherches ont proposé diverses taxonomies de questions. Travaux antérieurs sur la question La classification distingue deux types de taxonomie :à la taxonomie et hiérarchique taxonomie. Pendant que en taxonomie n'a qu'un seul niveau de classe sans sous-classes, taxonomie hiérarchique a des classes à plusieurs niveaux. Dans (Li et Roth, 2002), les auteurs proposent la taxonomie hiérarchique la plus populaire largement utilisée par la communauté SQR.Pour la langue arabe, des auteurs proposent une méthode pour

classer les questions arabes en à la taxonomie pour récupérer des réponses précises en utilisant des expressions régulières et des grammaires sans contexte. Dans [13], les auteurs proposent une à la taxonomie avec des classes de hauteur à l'aide de deux machines techniques d'apprentissage (SVM et multinomial Naïve Bayes) pour classer l'arabe question[23].

Faire un classificateur basé sur des règles pour la question arabe est une tâche très coûteuse en termes de temps et d'effort. Il a besoin d'une approche linguistique cohérente et utilise cartographie morphologique, très sensible à tout changement de vocabulaire et dépend de la taxonomie. Cependant, la classification de la question à l'aide de la machine les techniques d'apprentissage peuvent être facilement adaptées à une nouvelle taxonomie et peuvent être formées pour couvrir le plus grand vocabulaire.

Dans notre cas, nous avons adopté une plate taxonomie avec sept classes : Chose, Définition, Personne, Emplacement, Date Heure, Quantité et Description. La classe des questions détermine le type de réponse souhaité. La classe de réponse peut être vue dans Question Answering over Linked Data (QALD) comme plage du prédicat dans le triple contenant la réponse souhaitée. Par exemple, pour la question : « qui est le directeur du Parrain? هو مخرج العراب"?./man houa mokhrij al arab/, La réponse se trouve dans le triple <Parrain, réalisateur, Francis Ford Coppola>.Le directeur de prédicat a comme domaine lm et comme range person (<director, domaine, film> <directeur, portée, personne>). Ces informations peuvent aider le système à déterminer le prédicat dans l'ontologie et le type de la ressource demandée. Tableau 2 montre quelques questions arabes avec leurs mots interrogatifs correspondants et Des classes.

Table 2:Exemples de classes de questions et de noms interrogatifs

Question	Class	Interrogative noun
ما هي ديانة بليناس الحكيم ؟\maA hiya diyAnaħ balynAs Al.Hakym ?\ ما هي اليونسكو ؟\maA hiya Al.ywnis.kw ?\	Thing Definition	ما\maA\
من اكتشف الكهرباء ؟\man Ak.tašafa Al.kah.rabA' ?\ من هو تيم برنرز لي ؟\man huwa tym bar.nar.z liy ?\	Person	من\man\
أين تقع الصّحراء الكبرى ؟\Âay.na taqaçu AlS~aH.rA' Al.kub.raý ?\	Location	أين\Âay.na\
متى ولد مالك بن النّبي ؟\mataý wulida mAlik bn Aln~aby ?\	DateTime	متى\mataý\
كم عدد قتلى حرب الجزائر ؟\kam çadadu qat.laý Har.bu Al.jazaAŷir ?\	Quantity	كم\kam\
كيف قتل جورج واشنطن ؟\kay.fa qutila jwr.j wAšin.Tun ?\	Description	كيف\kay.fa\

Dans la tâche de classification des questions de notre système, nous utilisons des techniques d'apprentissage automatique afin de former un modèle qui classe la question arabe selon notre à la taxonomie contenant des classes de sept questions. (La figure 12) montre notre approche classe les questions dans le système global proposé.

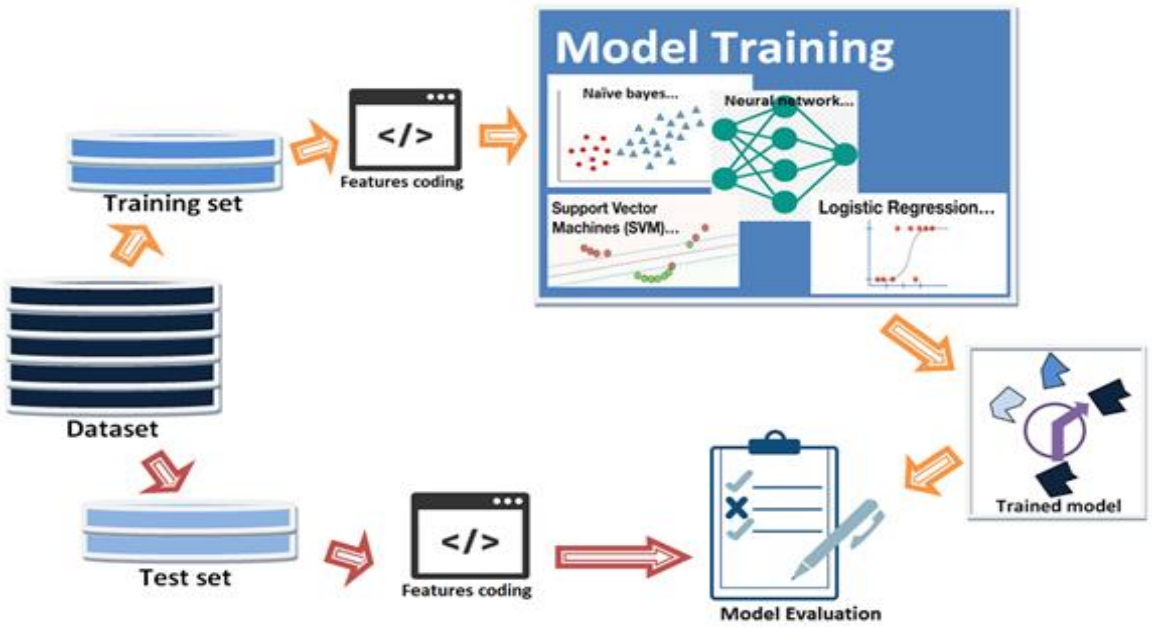

Figure 12:Approche de classification des questions basée sur des techniques d'apprentissage automatique.

Pour résoudre le problème de la classification des questions à l'aide d'un classificateur d'apprentissage automatique, nous avons besoin d'une représentation quantitative du texte d'entrée. Cette représentation est souvent appelé modèle vectoriel ou modèle caractéristique [15].

A ce stade, nous devons répondre à deux questions essentielles avant de poursuivre la description des autres composants du système : quelle technique de modèle de caractéristiques est mieux codifier une question ? Et quelle technique d'apprentissage automatique est la meilleure à classer une question?

Pour répondre à la première question, nous avons étudié trois modèles de caractéristiques : (1) Sac de mots : le modèle du sac de mots (BOW) est une représentation qui devient arbitraire texte en vecteurs de longueur fixe en comptant combien de fois chaque mot apparaît,(2) TF-IDF : TF-IDF signifie "Term Frequency - Inverse Document Frequency" modèle qui est une représentation qui transforme le texte en vecteurs de longueur fixe en quantifier un mot dans une question, et calculer un poids pour chaque mot qui signifie L'importance du mot dans la question et l'ensemble de données, et (3) Caractéristiques proposées :nous supposons que les informations essentielles dans une question sont le premier mot de la question et la structure morphologique de la question. Ainsi, le premier mot de la question, qui est, en général, un mot interrogatif pour l'arabe langue, est le signe le plus crucial de la question car il contient l'intention de la question. On utilise aussi le bigramme du mot interrogatif, le mot adjacent dans la question ainsi que la partie des balises vocales des trois premiers mots. Celles-ci sont les caractéristiques qui doivent être codées (processus de codage des caractéristiques) avant de classer une question[16].

7- Evaluation :

Pour répondre à la deuxième question, une expérience a été réalisée. Nous avons commencé avec 250 questions comme ensemble de données. Deux cents questions ont servi de formation défini et codé pour créer un modèle entraîné pour chaque technique d'apprentissage automatique

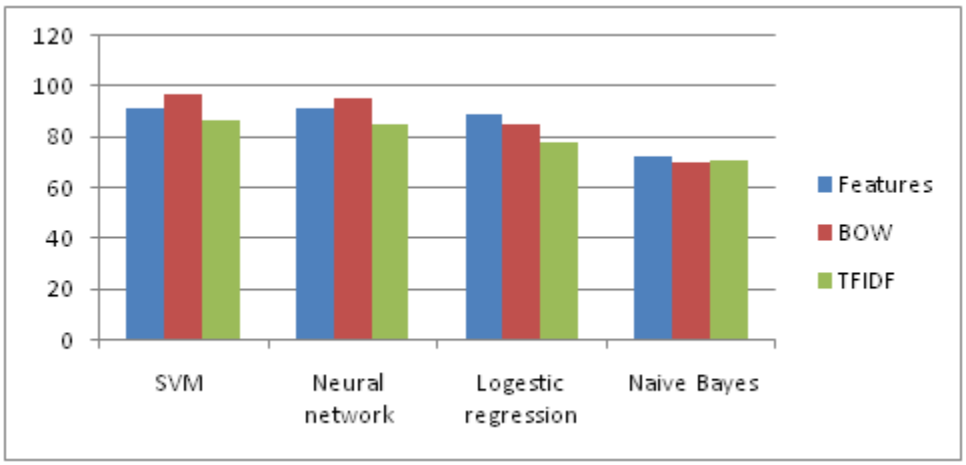

Figure 13:Le taux de réussite du processus de classification en utilisant différents apprentissage automatique technique.

(Machine à vecteurs de support (SVM), Réseau de neurones, Naive Bayes et Régression logistique).Ensuite, 50 questions ont été utilisées comme un ensemble de tests pour voir quelle technique donne le meilleur résultat de classement. Bien entendu, la classification s'est faite selon l'Adopté à la taxonomie (Chose, Définition, Personne, Emplacement, DateHeure, Description,quantité et description). La métrique utilisée pour comparer la machine learning techniques était le score de classification de précision qui renvoie le pourcentage de la questions correctement classées. Les résultats obtenus sont illustrés dans le graphique de (Figure 13).

Les résultats ci-dessus sont prometteurs et montrent la puissance des techniques d'apprentissage automatique lorsqu'il s'agit de systèmes de questions-réponses. Cette étude comparative montre que, dans notre système, la technique SVM, avec le modèle BOW pour codifier les questions, est le choix le plus efficace pour classer une question par rapport aux autres techniques d'apprentissage.

Maintenant, prenons la métrique F-mesure [22].pour évaluer le processus de classification séparément à l'aide de SVM. F-mesure est l'harmonique pondérée moyen de précision et de rappel. Il est défini comme suit : F-mesure = (2x Rappel x précision)/(Rappel + Précision), où Rappel est la proportion du correctement attribuer des questions à une classe de toutes les questions qui doivent être attribuées, et La précision correspond aux questions correctement attribuées à une classe de toutes les questions qui ont été attribués.

Les résultats de (la figure 13) montrent davantage que la SVM est la technique la plus efficace pour classer une question dans notre système.

8- Conclusion et perspectives :

La méthode de classification des questions en langue arabe ce base sur

L’apprentissage automatique et les règles NLP et se compose de quatre tâches : prétraitement, classification des questions, extraction de mots-clés et extraction de ressources.

En particulier, la tâche de classification des questions utilise des techniques d'apprentissage automatique (SVM, Réseau de neurones, Naïve Bayes et Régression logistique). Le formel le composant de requête utilise les sorties du module précédent pour déterminer le prédicat qui sera utilisé par le dernier composant, notamment le module de génération de réponses, pour définie la requête SPARQL qui nous donnera la bonne réponse.

Pour évaluer la tâche de classification des questions du module d'analyse des questions, nous a défini un ensemble de données de 250 questions. Cet ensemble de données a été divisé en deux sous-ensembles :

Le premier (200 questions) a été utilisé pour apprendre (former) des modèles de la machine techniques d'apprentissage; la seconde (50 questions) a été utilisée pour comparer ces techniques. L'étude comparative a montré que le SVM est le plus efficace technique pour classer les questions arabes dans notre système.

Les campagnes d'évaluation les plus populaires sur Question Answering over Linked Les données (QALD) ne fournissent pas encore d'ensemble de données arabe jusqu'au dernier défi QALD-9 [12]. Pour cette raison, nous avons construit nos ensembles de données d'évaluation.

ملخص :

في مذكرتنا قمنا بتصنيف الأسئلة، كما استخدمنا تقنيات التعلم الآلي لتدريب نموذج يصنف السؤال العربي وفقًا لتصنيفنا الذي يحتوي على فئات من سبعة أسئلة.

فنحتاج إلى تمثيل كمي لنص الإدخال. غالبًا ما يُطلق على هذا التمثيل اسم نموذج متجه أو نموذج مميز.

لذا يجب علينا أن نجيب على سؤالين أساسيين قبل الاستمرار في وصف المكونات الأخرى للنظام: ما هي تقنية نموذج الميزة الأفضل لتدوين سؤال؟ وما أسلوب التعلم الآلي الأفضل في ترتيب السؤال؟

بدأنا بـ 250 سؤالاً كمجموعة بيانات. مائتا سؤال أستخدم كتدريب محدد ومشفّر لإنشاء نموذج مدرب لكل أسلوب من تقنيات التعلم الآلي.

النتائج كانت واعدة وتظهر قوة تقنيات التعلم الآلي عندما يتعلق الأمر بأنظمة الأسئلة والأجوبة. ، و منه نستنتج الخيار الأكثر فعالية لتصنيف السؤال مقارنة بتقنيات التعلم الأخرى.

Référence et Bibliographie

[1] Abdelnasser H, Mohamed R, Ragab M, Mohamed A, Farouk B, El- Makky N, Torki M (2014)

[2] Al-Bayan: An Arabic Question Answering System for the Holy Quran.

[3] Ahmed W, Babu AP (2016) Question Analysis for Arabic Question Answering Systems. International Journal on Natural Language Computing (IJNLC) 5(6)

[4] Al Chalabi HM, Ray SK, Shaalan K (2015) Question classification for Arabic question answering systems.

[5] In: 2015 International Conference on Information and Communication Technology Research (ICTRC), IEEE, pp 310{313

[6] Al-Harbi O (2019) A Comparative Study of Feature Selection Methods for Dialectal

[7] Arabic Sentiment Classification Using Support Vector Machine. arXiv preprint arXiv:190206242

[8] Al-Zoghby AM, Ahmed ASE, Hamza TT (2013) Arabic semantic web applications: a survey. Journal of Emerging Technologies in Web Intelligence 5(1):52{69

[9] AlAgha I (2015) Using Linguistic Analysis to Translate Arabic Natural Language Queries to SPARQL. International Journal of Web and Semantic Technology (IJWesT) 6(3):29{35

[10] AlAgha I, Abu-Taha A (2015) AR2SPARQL: An Arabic Natural Language Interface for the Semantic Web. International Journal of Computer Applications 125(6)

[11] Albarghothi A (2018) An Ontology-based Semantic Web for Arabic Question Answering: The Case of E-Government Services. PhD thesis

[12] Albarghothi A, Khater F, Shaalan K (2017) Arabic question answering using ontology.

[13] Procedia Computer Science 117:183{191

[14] Atzori M, Mazzeo GM, Zaniolo C (2019) QA3: A natural language approach to question answering over RDF data cubes. Semantic Web (Preprint):1{18 Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval. ACM press New York

[15] Bekhti S, Al-Harbi M (2013) AQuASys: A Question-Answering System For Arabic.

[16] In: Proceedings of the 13th International Conference on Applied Computer Science (ACS '13), Proceedings of the 2nd International Conference on Digital Services, Internet and Applications (DSIA'13), WSEAS Press, Morioka City, Iwate, Japan

[17] Bizer C, Heath T, Berners-Lee T (2011) Linked data: The story so far, IGI Global, pp 205{227

[18] Bouziane A, Bouchiha D, Doumi N, Malki M (2018) Toward an Arabic Question Answering System over Linked Data.

[19] Jordanian Journal of Computers and Information Technology (JJCIT) Vol. 04(No. 02)

[20] DBpedia (2012) Dbpedia Mappings

[21] Diefenbach D, Singh K, Maret P (2018) Wdaqua-core1: a question answering service for rdf knowledge bases. In: Companion Proceedings of the The Web Conference 2018, International World Wide Web Conférences Steering Committee, pp 1087{1091 Green BF, Wolf AK, Chomsky C, Laughery K (1961)

[22] BASEBALL: An automatic question answering. In: Proceedings Western Joint Computer Conference, McGraw-Hill, vol 19, pp 207{216

[23] Habash N (2010) Introduction to Arabic natural language processing. Synthesis