N° d'ordre :

REPUBLIQUE ALGERIENNE DEMOCRATIQUE & POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR & DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DJILLALI LIABES
FACULTE DES SCIENCES EXACTES
SIDI BEL ABBES

THESE
DE DOCTORAT

Présentée par

Bouziane Abdelghani

Spécialité : Informatique
Option : Interopérabilité et intégration des systèmes d'informations dans le Web

Intitulée

Exploitation des données liées : Système Question-Réponse

Soutenue le 03 Avril 2019
Devant le jury composé de :

Président :(Sofiane BOUKLI HACENE , MCA , UDL-SBA)
Examinateurs : Maamar KHATER MCA Université de Saida
Reda ADJOUDJ, MCA, UDL-SBA
Moussa ALI CHERIF MCA UDL-SBA
Directeur de thèse : Djelloul Bouchiha Professeur CU-Naama
Co-Directeur de thèse : Mimoun MALKI Professeur UDL-SBA

Année universitaire 2018/2019

DÉDICACES

A MES CHERS PARENTS AVEC TOUT MON AMOUR

VOS PRIÈRES ET VOS BÉNÉDICTIONS M'ONT ÉTÉ D'UN GRAND SECOURS POUR MENER À BIEN MES ÉTUDES.

UNE DÉDICACE À MA FEMME POUR SON SOUTIEN. QUE DIEU RÉUNISSE NOS CHEMINS POUR UN LONG COMMUN SEREIN.

A MES FRÈRES HABIBE, ABDELKADER, MOKHTAT, BRAHIME, FARHAT ET HADI, MERCI D'AVOIR CRU EN MOI, MERCI DE VOTRE PATIENCE ET DE L'AFFECTION QUE VOUS M'AVEZ MANIFESTÉE DURANT CES ANNÉES. QUE CE TRAVAIL SOIT TÉMOIGNAGE DE MA RECONNAISSANCE

A MES CHÈRES SŒUR BAHRIA, GHANIA, TORKIA ET NORA, VOUS AVEZ TOUJOURS ÉTÉ PRÉSENTS PAR VOTRE AMOUR. VOTRE AFFECTION ET VOTRE SOUTIEN M'ONT ÉTÉ D'UN GRAND SECOURS AU LONG DE MA VIE PROFESSIONNELLE ET PERSONNELLE. VEUILLEZ TROUVER DANS CE MODESTE TRAVAIL MA RECONNAISSANCE POUR TOUS VOS EFFORTS.

REMERCIEMENTS

Je tiens à remercier Monsieur *Djelloul BOUCHIHA*, Professeur au Centre Universitaire de Naâma, qui m'a encadré tout au long de cette thèse et qui m'a fait partager ses brillantes intuitions. Qu'il soit aussi remercié pour sa gentillesse, sa disponibilité permanente, son sens de partage, ses bonnes mœurs et son esprit de responsabilité qu'ils mon fortement inspiré. Sa compétence, sa rigueur scientifique et sa clairvoyance m'ont beaucoup appris.

Je remercie aussi Monsieur Mimoun MALKI, Ecole Supérieure d'Informatique de Sidi Bel-Abbes, ESI-SBA, pour la confiance qu'il m'a accordée en acceptant la codirection de ma thèse.

Je souhaiterais adresser mes remerciements les plus sincères à Monsieur Nourddinne DOUMI pour sa collaboration scientifique, ces conseils, sa compétence, son partenariat et son amitié.

Je souhaiterais adresser mes remerciements les plus sincères à Monsieur Sofiane BOUKLI HACENE docteur à l'UDL de Sidi Bel Abbes, pour m'avoir fait l'honneur d'être président de jury. Je voudrai également remercier tous les membres du laboratoire EEDIS "Evolutionary Engineering and Distributed Information Systems Laboratory ", et également tous les professeurs, chercheurs et professionnels de ce laboratoire que j'ai connus.

Je remercie les examinateurs de ce travail, Monsieur Reda ADJOUDJ, docteur à l'Université de Sidi Bel Abbes, Monsieur Maamar KHATER, docteur à l'Université de Saida et Monsieur Moussa ALI CHERIF, docteur à l'université de Sidi Bel Abbes pour avoir accepté de faire partie de mon jury de thèse. Je les remercie pour leurs précieux commentaires et recommandations.

Table des matières

Liste des Figures

Liste des Tableaux

Liste des Listings

Chapitre 1 : Introduction générale

1. Contexte

Chaque jour, nous créons 2,5 quintillions ($2,5 * 10^{30}$) d'octets de données. Pour mettre cela en perspective, 90% des données dans le monde ont été créées au cours des trois dernières années seulement (Cloud 217), et avec l'apparition de nouveaux appareils, capteurs et technologies, le taux de croissance des données va probablement accélérer encore plus.

Cette augmentation rapide de stockage massif d'informations, et la popularité de l'utilisation du Web, permettent aux chercheurs de stocker des données et de les rendre accessibles au public. Toutefois, l'exploration de cette grande quantité de données fait de la recherche d'information une tâche complexe et coûteuse en termes de temps. Cette difficulté a motivé le développement de nouveaux outils de recherche adaptés, tels que les Systèmes Question-réponse (SQR).

Les Systèmes question-réponse (SQR), le meilleur scénario :

En fait, ce type de système permet à l'utilisateur de poser une question en langage naturel (LN) et de retourner la bonne réponse à sa question au lieu d'un ensemble de documents jugés pertinents, comme c'est le cas pour les moteurs de recherche. Le contexte dans lequel un système question-réponse est utilisé, c.-à-d., l'utilisateur anticipé, le type de questions, le type de réponses attendues et le format dans lequel les informations disponibles sont stockées, déterminent la conception du système. On peut distinguer deux catégories de systèmes question-réponse. Premièrement, les systèmes qui tentent de répondre à une question en accédant à des informations structurées contenues dans une base de données, une base de connaissances ou une ontologie. Dans cette catégorie, on distingue deux types de systèmes : les interfaces de langage naturel pour les bases de données et les systèmes question-réponse sur les données liées. Deuxièmement, les systèmes qui tentent à répondre à une question en analysant des informations non-structurées, telles que les textes clairs (document, page HTML, etc.) (Bouziane et al. 2015). Bien entendu, de nombreux systèmes sont hybrides, i.e., possédant les caractéristiques des deux types.

Pour le WEB 3.0, une nouvelle génération des SQR :

La plupart des SQR Web qui existent actuellement traitent les documents. La structure des informations requises sur le Web de documents affecte la précision de ces systèmes. Ces systèmes doivent interagir avec des bases de connaissances structurées et valides. Actuellement, la mutation de ces systèmes vers le Web de données semble nécessaire pour trouver des réponses correctes et précises aux questions. Les nouveaux systèmes question-réponse doivent interagir avec les données liées plutôt que les documents liés. Les données liées ont été discutées dans le cadre de la technologie du Web sémantique (Berners-Lee et al. 2006). Les données liées fournissent un paradigme de publication dans lequel non seulement les documents, mais aussi les données, peuvent constituer un citoyen de première classe dans le Web (Bizer et al. 2011). L'extension du Web actuel vers le Web sémantique nécessite la création de nouveaux systèmes question-réponse sur les données liées capable d'interagir avec le Web 3.0.

L'arabisation de la technologie, un besoin qui s'est fait ressentir :

L'arabe est la langue la plus parlée dans le groupe sémitique et la langue officielle ou co-officielle de 26 pays, parlée par plus de 422 millions de personnes au Moyen-Orient et en Afrique du Nord. L'arabe est la quatrième langue sur Internet avec 219 millions d'utilisateurs (MINIWATTS 2018). 54% des recherches Google au Moyen-Orient et en Afrique du Nord (MENA) sont désormais effectuées en arabe, 34% en anglais et 8% en français (Bas 2018).

2. Problématique

Il existe des ressources et des outils limités pour le traitement automatique de la langue arabe par rapport à la langue anglaise. L'autre difficulté est le manque de prise en charge de la langue arabe dans les technologies du Web sémantique. Le Web sémantique arabe est très loin des meilleures performances par rapport au Web sémantique anglais. Les difficultés du traitement du langage naturel arabe entravent le développement du Web sémantique arabe, car le TAL arabe est une composante importante pour l'arabisation du Web sémantique.

La mise en place d'un système question-réponse pour le Web sémantique arabe doit donc confronter plusieurs obstacles selon trois dimensions à savoir :

Système :

Les systèmes question-réponse ont tendance à être assez complexes, avec plusieurs modules. Chacun de ces composants et chacune des techniques qu'ils utilisent ont un impact certain sur les performances globales d'un système question-réponse. Il est donc très difficile d'étudier tous les aspects d'un système question-réponse. Par conséquent, certaines limites doivent être soulevées.

Sémantique :

Comme nous visons dans notre thèse à exploiter des données liées dans le Web sémantique, nous devons faire face à des données semi-structurées, non encore largement adoptées dans le Web. Les données les plus répandues dans le Web sont exprimées principalement en HTML. C'est pourquoi, il n'y a que peu de travaux qui ont traité ce problème. Quand elles existent, les données liées ne peuvent être exploitées que par un langage dédié, à savoir SPARQL.

Langue :

La langue arabe se caractérise par une morphologie relativement complexe. Elle possède un système riche d'inflexion morphologique. L'arabe a également un haut degré d'ambiguïté résultant de son système d'écriture optionnel diacritique, de son écart courant par rapport aux normes orthographiques, et de l'absence de majuscules (Bouziane et al. 2018).

Les outils et ressources existants traitent principalement de l'arabe standard moderne (MSA), mais le monde arabe utilise différents dialectes arabes, tels que l'Arabe égyptien (EGY), levantin (LEV), golfe arabe (GLF), nord-africain (maghrébin) Arabe (Mag), Arabe irakien (IRQ), Arabe yéménite (Yem) (Guo and Ren 2009).

3. Objectif du travail

Dans cette thèse, nous visons à créer un système question-réponse sur le Web sémantique pour la langue arabe. Ce défi nous mène à faire face, d'une part à la difficulté de réduire la complexité du langage naturel par rapport au langage de requête SPARQL, et de l'autre part à des difficultés liées à la langue arabe, son traitement automatique et sa compatibilité avec le Web sémantique.

Le système proposé est composé de trois modules consécutifs. Chacun est composé de plusieurs ‘, Premièrement, le système reçoit en entrée une question exprimée en langage naturel arabe, à partir de laquelle il fournit la ressource et les mots-clés en utilisant des techniques de TAL dans le premier module. Ensuite, dans un deuxième module, il utilise une requête SPARQL pour explorer l'ontologie et produit le prédicat exact en comparant les termes entre mots-clés et prédicats. Enfin, dans un dernier module, notre système fournit une réponse à la question de l'utilisateur en exécutant une requête SPARQL finale appropriée.

4. Principales contributions

Dans cette thèse, nous proposons un système question-réponse capable d'exploiter les données liées en langue arabe.

Motivés par cet objectif :

- ☞ nous donnons une classification basée en particulier sur les dimensions de la langue et de la structure des données,
- ☞ nous introduisons les concepts du Web sémantique et des données liées,
- ☞ nous expliquons la position de la langue arabe vis-à-vis des technologies actuelles du Web sémantique,
- ☞ nous citons les systèmes question-réponse, et nous analysons leurs propositions selon différents points de vue,
- ☞ nous actualisons les états de l'art existants en ajoutant des travaux récents,
- ☞ nous présentons des statistiques par des histogrammes graphiques qui donnent une vision claire aux chercheurs travaillant dans ce domaine,
- ☞ nous présentons un système question-réponse pour exploiter les données liées arabes,
- ☞ et nous évaluons et discutons les résultats de notre système pour montrer son efficacité et penser à son perfectionnement.

5. Organisation de la thèse

Le reste du mémoire est organisé comme suit :

Chapitre 2 : dans ce chapitre, nous discutons certaines approches et concepts des SQR, allant des approches antérieures au plus récentes. Si l'on se penche sur les approches antérieures, ce n'est pas seulement une valeur historique, mais elle révèle également des problèmes généraux en SQR et la manière dont ces problèmes ont été abordés au fil des années. L'objectif de ce chapitre est de synthétiser et d'identifier les principales notions relatives aux systèmes question-réponse en tenant compte d'un certain nombre d'approches et en discutant l'évaluation des SQR.

Chapitre 3 : dans ce chapitre on introduit un background sur deux axes de recherches liés à notre problématique qui sont le Web sémantique et le TAL Arabe. On introduit les notions de base sur la technologie du Web sémantique, on présente les définitions de base sur la langue arabe et les outils du TAL arabe, on étudie la situation actuelle du Web sémantique arabe, et on présente un exemple du projet DBpedia, à savoir "*le chapitre arabe du DBpedia*".

Chapitre 4 : ce chapitre présente un état de l'art sur les SQR. On présente une vue détaillée sur les trois types de SQR, leurs caractéristiques, approches, architecture, modules, et évaluations. Les évaluations et les statistiques donnent une vision plus précise et plus crédible sur les performances des SQR.

Chapitre 5 : ce chapitre présente un nouveau système question-réponse pour le Web sémantique arabe. L'approche discutée dans ce chapitre est une première version d'un nouveau système question-réponse en langue arabe, indépendant du domaine, sur les données liées, qui vise en particulier à aider les utilisateurs arabes à explorer le Web sémantique basé sur une ontologie arabe. Suite à la présentation de l'architecture des composantes de traitement nécessaires pour sa réalisation et son évaluation, nous revenons sur le contexte du travail et sur les manières d'exploiter les résultats que nous présentons dans cette thèse.

Chapitre 6 : on clôture cette thèse par une synthèse concluant nos principales contributions dans le domaine des SQR pour le Web sémantique arabe. Ce chapitre présente également les perspectives liées à la poursuite de ce travail ainsi qu'aux nouveaux thèmes de recherche qui nous paraissent les plus pertinents.

Chapitre 2 : Les Systèmes Question-Réponse

1. Introduction

Les systèmes question-réponse (SQR) est un domaine de recherche en pleine croissance qui regroupe des recherches issues de la recherche d'information (IR), de l'extraction d'informations (EI), du Web sémantique, et du traitement automatique du langage naturel (TAL). Les techniques et méthodes développées à partir des systèmes question-réponse inspirent de nouvelles idées dans de nombreux domaines étroitement liés, tels que la recherche de documents, la reconnaissance des entités nommées (REN), la génération des ontologies, les données liées, etc.

La construction des systèmes question-réponse a commencé depuis la tentative faite par Green et ses collaborateurs (1961) à travers le système appelé «BASEBALL» (Green et al. 1961). En 1965, l'article de Simmons (1965) (Simmons 1965) traitait des efforts déployés par quinze systèmes question-réponse pour répondre automatiquement aux questions anglaises. Ces premiers systèmes implémentés se sont concentrés sur des domaines et des questions spécifiques. Ensuite, le domaine a évolué vers de nouvelles tendances grâce à la disponibilité des informations en ligne, aux séries de conférences d'évaluation organisées et aux tâches des systèmes question-réponse. Les sociétés les plus importantes, telles que Google, Yahoo, Microsoft et IBM, se sont intéressées à la réalisation de tels projets, ce qui témoigne la popularité croissante de ce secteur.

En règle générale, les systèmes question-réponse sont construits autour d'une architecture générale qui combine plusieurs composantes dans le but de répondre automatiquement aux différents types de questions des utilisateurs. L'efficacité de ces systèmes est mesurée au moyen d'un processus d'évaluation utilisant des ensembles de tests pertinents liés à la langue cible.

L'objectif de ce chapitre est d'introduire les principaux concepts des systèmes question-réponse. Le reste de ce chapitre est organisé comme suit : la section suivante présente l'historique et l'évolution chronologique des systèmes question-réponse. La section 3 décrit l'architecture standard du système question-réponse. La section 4 présente une classification des systèmes question-réponse existants selon plusieurs critères. La section 5 présente les principales approches pour les systèmes question-réponse. La section 6 fournit ensuite des informations sur les campagnes d'évaluation des systèmes question-réponse TREC, CLEF et QUALD. Le chapitre se termine par une conclusion dans la section 7.

2. Historique des SQR

L'évolution de l'informatique dans les années 1940 et 1950 a donné naissance aux premiers véritables ordinateurs qui sont Turing complets, électroniques et relativement rapides. L'intelligence artificielle (IA) voit le jour quelques années plus tard en 1956 à l'occasion de l'école d'été organisée à l'université de Dartmouth. L'IA a fixé comme objectif d'imité le comportement humain, notamment pour ce qui touche au raisonnement, aux mathématiques, à la perception et la compréhension du langage. Cette dernière à donner naissances aux premiers travaux en Traitement Automatique des Langues (TAL).

Les premiers travaux en TAL se sont focalisés sur le développement des premiers traducteurs automatique qui sont basiques, traduisant des phrases simples de la langue russe

vers l'anglais en 1954. En 1962, la première conférence sur la traduction automatique est organisée à MIT. Le développement des recherches dans le domaine du TAL a contribué à la création des premiers systèmes question-réponse.

L'histoire des systèmes question-réponse remonte aux années soixante. Le premier état de l'art sur les systèmes question-réponse publié en 1965 dans lequel plusieurs systèmes on été étudiés pour l'anglais sur les cinq années précédentes (Simmons 1965). En 1978, le premier ouvrage classique sur les systèmes question-réponse a été publié. Il est basé sur les travaux de Lehnert's (Lehnert 1977) dans lesquels il a proposé un système QR basé sur la sémantique et le raisonnement.

Les premiers SQR qui ont vu le jour même avant 1978, sont des interfaces de langage naturel pour les bases de données. Ces systèmes supportent d'une façon minime le langage naturel et sont adaptés à un domaine spécifique et restreint. La solution offerte par ces systèmes est la translation des questions écrites en langage naturel en un langage de requête pour les bases de données comme SQL. Les ILNBD les plus connus de l'époque sont LUNAR (Woods et al. 1972) et BASEBALL (Green et al. 1961).

SHRDLU (WINOGRAD 1971) est un système question-réponse capable de mener un simple dialogue avec un utilisateur humain sur un petit monde d'objet. SHRDLU est une application de TAL écrit par Terry Winograd au MIT (Laboratoire d'intelligence artificielle) en 1968-71.

L'exemple suivant présente une conversation entre une personne et le système SHRDLU.

```
Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I don't understand which pyramid you mean.
Person (changing their mind): Find a block which is taller than
the one you are holding and put it into the box.
Computer: By "it", I assume you mean the block which is taller
than the one I am holding.
Computer: OK.
Person: What does the box contain?
Computer: The blue pyramid and the blue block.
Person: What is the pyramid supported by?
Computer: The box.
```

Listing 1. Conversation entre une personne et SHRDLU

Dans les années 70-80, la recherche dans la linguistique computationnelle et le développement de la théorie compréhensive ont donné lieu à des projets ambitieux dans le domaine des systèmes question-réponse et la compréhension automatique du texte. L'UNIX Consultat (UC) est un exemple d'un système QR développé par Robert Wilensky à Berkeley dans la fin des années 80 (Robert et al. 1988). Ce système est doté d'une base de connaissances dans son domaine. Un autre exemple de la compréhension du texte est le

système LILOG qui opère sur le domaine du tourisme en Allemagne. Le système ILLOG et celui de Robert Willensky n'ont pas dépassé le stade de la simple démonstration, mais ils ont contribué au développement des théories dans la linguistique computationnelle et le raisonnement automatique.

Pendant les années 90, les nouvelles technologies de l'information et de la communication et du World Wide Web ont largement contribué au développement des systèmes question-réponse appliqués au domaine ouvert (QRDO). Les données non structurées largement disponibles dans le Web ont augmenté le besoin de créer des nouveaux systèmes QR pour exploiter les données textuelles. START (Katz 1993) est le premier système QR orienté Web développé par Bris Kats en décembre 1993. Avec l'intérêt renaissant accordé aux systèmes QR, un challenge des systèmes question-réponse appliqués dans un domaine ouvert apparaît au sein de la campagne d'évaluation Text Retrieval Conference TREC du National Institute of Standards and Technology (NIST) et U.S. Department of Defence initiée en 1992. La première édition de la tâche QR (« QA track ») a eu lieu en 1999 (Voorhees and Tice 1999). Cet espace offre un cadre d'évaluation commun au développement des systèmes QR et marque l'essor des systèmes question-réponse au domaine ouvert (QRDO).

Fin des années 90 était l'ère des moteurs de recherche en incubation. Au milieu de l'année 2000, la recherche est devenue maturée et s'est focalisée sur la recherche orientée Web. Divers systèmes question-réponse ont été développés. La succession des évaluations annuelles des systèmes question-réponse, telles que Text Retrieval Conference « TREC » (Voorhees and Harman 2017), Conference and Labs of the Evaluation Forum « CLEF » a permis l'amélioration de la performance à un degré avancé des SQR.

Récemment, avec l'émergence du Web sémantique, de plus en plus de données structurées en Resource Description Framework (RDF) sont publiées sur le Web. Le défit de savoir comment les utilisateurs du Web peuvent accéder à ce corpus de connaissances devient d'une importance cruciale. Jusqu'à présent, il n'y a pas beaucoup d'outils qui permettraient aux utilisateurs finaux d'exploiter la puissance du Web sémantique tout en cachant la complexité derrière une interface intuitive et facile à utiliser. Un défi important pour le Web sémantique, mais aussi pour la communauté de traitement du langage naturel, est de mettre à l'échelle les SQR selon cette nouvelle forme de données.

QALD (Question Answering system over Linked Data) lancé en 2011, est une série de compagnes d'évaluation pour les systèmes question-réponse sur les données liées. Jusqu'à présent, ESWC «European Semantic Web Conference» a été organisé en tant qu'atelier dans le cadre du laboratoire Question Answering à CLEF. SWIP est un exemple d'un SQR sur les données liées.

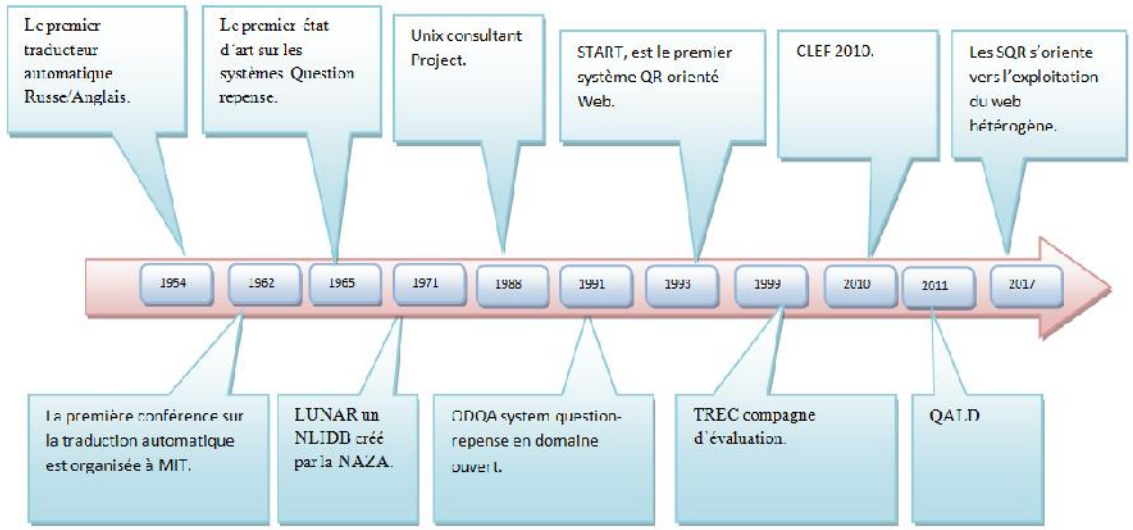

Figure 1: Chronologie de développement des systèmes question-réponse

3. Architecture générale

Selon la littérature, un SQR est composé au minimum de trois modules : Analyseur de la question, Sélectionneur du document /exploitation de la base, Extraction de la réponse. D'autres modules comme l'Expansion de la requête, Enrichissement de la question originale avec des termes connexes et Validation de la réponse, peuvent être ajoutés dans l'architecture des SQR de la Figure 2.

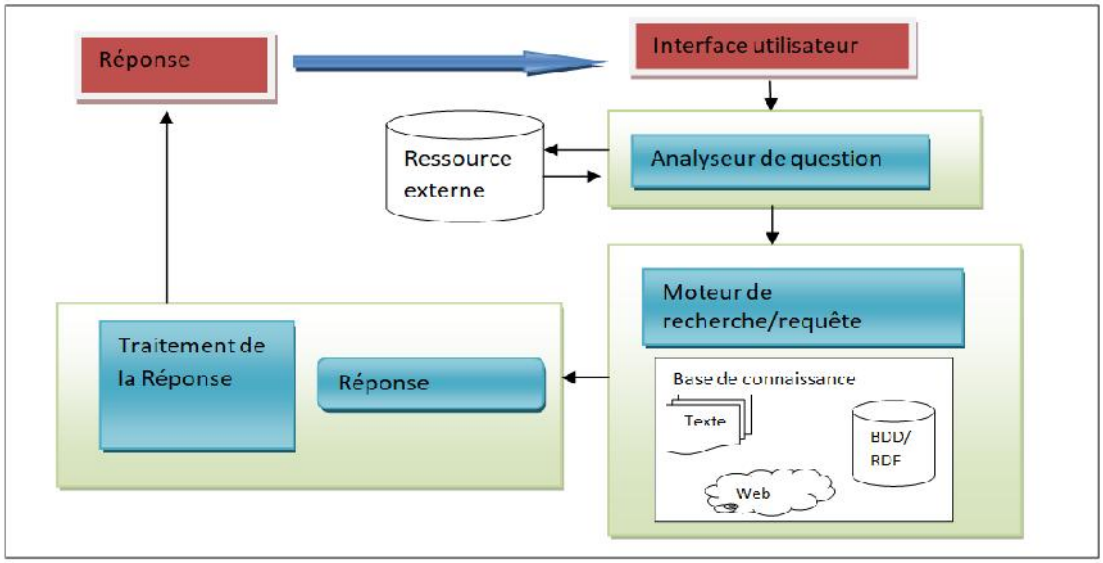

Figure 2: Architecture globale d'un SQR

L'utilisateur pose une question en langage naturel via une interface graphique. L'analyseur de question fait des traitements en utilisant des ressources externes en utilisant des approches linguistiques et statistiques : morphosyntaxique, sémantique, classification des questions, reconnaissance des entités nommées NER, détections des classes et des mots clés de la question d'utilisateur. Ensuite, la question sera représentée sous formes logique, table de

BDD, langage de requête ou un ensemble de mots clés. La base de connaissances sera interrogée par la requête formulée dans l'étape précédente sous une forme adéquate par rapport à la nature des données structurées ou non-structurées. L'exécution de la requête par le moteur de recherche pour les données textuelles ou par le moteur de requête pour les données structurées, produit une réponse sous forme d'un morceau de texte, document, table BDD ou de fichier RDF qui sera par la suite traité et vérifié pour produire la réponse exacte.

4. Classification des SQR

La classification la plus complète des SQR a été proposée par Mishra (Mishra and Jain 2016). Mishra identifie huit critères de classification des SQR existants. Ces critères sont :

4.1. Domaine d'application

Le défi de générer des réponses pour des questions est relatif aux types de questions posées (Moldovan et al. 2000; Burger et al. 2001). Certains utilisateurs peuvent avoir besoin d'informations générales sur un sujet général; d'autres peuvent exiger des informations spécifiques d'un domaine d'application particulier. Par conséquent, la sélection du domaine comme base de classification des SQR peut être un choix naturel.

4.2. Type de question

Le défi de générer des réponses aux questions des utilisateurs est directement lié au type de la question posée. Ainsi, la classification des questions dans un SQR affecte directement la précision du system. Les résultats montrent que 36,4% des erreurs sont générées en raison d'une mauvaise classification des questions posées dans les SQR (Moldovan et al. 2003).

Dans (Mishra and Jain 2016), les auteurs classifient les SQR en fonction des types de questions posées par les utilisateurs. Les différentes catégories sont :

i. questions de type factoid.
ii. des questions de type liste.
iii. des questions de type hypothétique.
iv. questions de confirmation.
v. questions causales.

4.3. Type d'analyse effectuée sur la question

Mishra et Jain (Mishra and Jain 2016) classifient les SQR en fonction des types d'analyses effectuées sur les questions. Les différentes catégories sont :

i. analyse morphologique.
ii. analyse syntaxique.
iii. analyse sémantique.
iv. analyse pragmatique et discursive.
v. analyse du type de réponse attendue.
vi. concentrer la reconnaissance des questions.

4.4. Type de source de données

C'est une classification des SQR en fonction des types de données présentés dans la base à interroger. Les différentes catégories sont :

i. source de données structurée,
ii. source de données non-structurée.

Les deux catégories citées ci-dessus sont détaillées dans le chapitre 4 : état de l'art sur les SQR.

4.5. Type de la fonction d'appariement utilisée dans différents modèles de récupération

Dans (Mishra and Jain 2016), les auteurs classifient les SQR en fonction des types de fonctions de correspondance utilisées dans différents modèles de récupération. Les différentes catégories sont :

i. définir des modèles théoriques.
ii. modèles algébriques.
iii. modèles de probabilité.
iv. modèles basés sur les caractéristiques.
v. analyse du type de réponse attendue.
vi. modèles basés sur un graphe conceptuel.

4.6. Caractéristique des sources de données

(Mishra and Jain 2016) Classifie les SQR en fonction des caractéristiques des sources de données.

Les différentes catégories sont:

i. la taille de la source.
ii. la langue.
iii. hétérogénéité.
iv. genre.
v. médias.

4.7. Forme de la réponse

C'est une classification basée sur des formes de réponses générées par les SQR. Les différentes catégories sont:

i. réponse extraite.
ii. réponse générée.

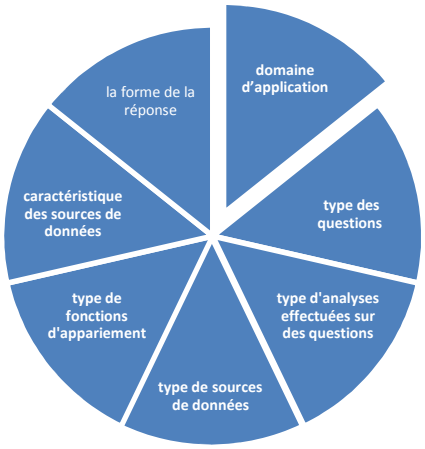

Figure 3: Classification des SQR

5. Les approches des SQR

Les SQR combinent des techniques issues de l'intelligence artificielle, du traitement automatique du langage naturel, de l'analyse statistique, de l'appariement de modèles, de la recherche d'information et de l'extraction d'information. La plupart des travaux récents intègrent une partie ou la totalité de ces approches pour construire des systèmes performants capables de faire face aux faiblesses de ces approches.

Dans la littérature, on distingue trois grandes approches :

1. Approche linguistique.
2. Approche statistique.
3. Approche de filtrage par pattern (motif).
   - ☞ Basée sur la surface par pattern.
   - ☞ Basée sur les modèles.

5.1. Approche linguistique

Un SQR nécessite la compréhension du texte en langage naturel, la linguistique et des connaissances générales. Par conséquent, beaucoup de chercheurs utilisaient des méthodes basées sur l'IA (intelligence artificielle) qui intègrent des techniques de TAL (traitement automatique du langage naturel) et une base de connaissances ou un corpus pour construire des modules pour SQR. Les connaissances sont organisées sous la forme de règles de production, de logiques, de trames, RDF (représenté avec des relations triplets), d'ontologies et de réseaux sémantiques, qui sont utilisés lors de l'analyse de la paire question-réponse. Des techniques linguistiques, telles que la Tokenization, le POS tagging et l'analyse syntaxique, sont implémentées sur les modules de traitement de la question de l'utilisateur pour formuler

une requête précise qui extrait simplement la réponse correspondante de la base de données structurée.

Cependant, le déploiement d'une base de connaissances sur un domaine spécifique pose un problème de portabilité, car un domaine d'application différent requiert des règles de grammaire et des règles de mapping différentes. De plus, la construction d'une base de connaissances appropriée est un processus qui prend beaucoup de temps, de sorte que ces systèmes sont généralement appliqués à des problèmes ayant des besoins d'information à long terme pour un domaine particulier.

Certains SQR existants exploitent le Web en tant que ressource de données. Ces systèmes appliquent leurs propres heuristiques pour stocker des informations à partir du Web (données structurées ou non-structurées) dans la base de connaissances locale, à laquelle il faut ensuite accéder et s'appuyer sur des techniques linguistiques et de filtrage pour la génération des réponses à partir des documents pour les données non-structurées ou à partir du résultat d'exécution des requêtes pour les données-structurées.

5.2. Approche statistique

Actuellement, la croissance rapide des données Web disponibles (structurées et non-structurées) a accru l'importance des approches statistiques. Ces approches mettent en avant de telles techniques, qui peuvent non seulement traiter la très grande quantité de données mais aussi leur hétérogénéité.

De plus, les approches statistiques peuvent formuler des requêtes en langage naturel. Ces approches nécessitent fondamentalement une quantité suffisante de données pour un apprentissage statistique précis, mais une fois correctement apprises, elles produisent de meilleurs résultats par rapport à d'autres approches informatiques.

L'algorithme d'apprentissage statistique peut être facilement adapté à un nouveau domaine indépendamment de toute forme de langage. Cependant, l'un des inconvénients majeurs des approches statistiques est qu'elles traitent chaque terme indépendamment et ne parviennent pas à identifier les caractéristiques linguistiques pour la combinaison de mots ou de phrases.

En général, les techniques statistiques ont jusqu'ici été appliquées avec succès aux différentes étapes d'un système QR. Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais support vector machine, SVM), les classificateurs bayésiens et les modèles à entropie maximale, sont des techniques qui ont été utilisées à des fins de classification de questions. Ces mesures statistiques analysent les questions permettant de prédire le type de la réponse attendue par les utilisateurs. Ces modèles sont entrainés sur un corpus de questions ou de documents qui a été annoté avec les catégories mentionnées dans le système.

L'un des travaux pionniers basés sur le modèle statistique était le SQR statistique d'IBM (Ittycheriah et al. 2000). Ce système utilise le modèle d'entropie maximale pour la classification questions/réponses en fonction de plusieurs caractéristiques de N-gramme ou des ensembles de mots.

5.3. Approche de filtrage par pattern

Cette approche utilise la puissance expressive des patterns de texte pour remplacer le traitement sophistiqué dans d'autres approches informatiques. Les patterns sont des formes de langage ; la reconnaissance des patterns est étudiée dans de nombreux domaines, notamment la psychologie, la psychiatrie, l'ethnologie, les sciences cognitives et l'informatique (Kocaleva et al. 2016). Par exemple, la question «Où se tenait la coupe du monde de cricket 2012?» suit le pattern «Où se tenait <Nom de l'événement>? et la réponse suit le pattern "<Nom de l'événement> a eu lieu à <Lieu>". Actuellement, de nombreux SQR apprennent automatiquement les patterns de texte à partir de passages de texte plutôt que d'utiliser des connaissances ou des outils linguistiques compliqués, tels que l'analyse syntaxique, l'identificateur d'entité nommée NER, l'ontologie, WordNet, etc., pour récupérer des réponses.

La simplicité de tels systèmes les rend plutôt favorables aux petites et moyennes applications, qui ne nécessitent pas des solutions complexes demandant beaucoup de temps et de compétences pour installer et maintenir le système.

5.4. La surface par pattern

Cette approche extrait les réponses de la structure linguistique des documents trouvés par le moteur de recherche en s'appuyant sur une longue liste de patterns. La réponse à une question est identifiée sur la base de la similarité entre les patterns ayant une certaine sémantique. Ces patterns sont comme des expressions régulières. Bien que la conception d'un tel ensemble de patterns nécessite beaucoup de compétences humaines et de temps, mais l'approche a montré une grande précision aussi.

Initialement, la méthode basée sur la surface par pattern vise à trouver des réponses à des questions factuelles, car leurs réponses sont limitées à une ou deux phrases. Afin de concevoir un ensemble optimal de patterns, la majorité des SQR basés sur la méthode de surface par pattern utilisent la méthode décrite par Ravichandran et Hovy (Ravichandran and Hovy 2002). Ils ont mis en place une méthode d'apprentissage automatique qui utilise le bootstraping pour construire un grand ensemble de patterns commençant seulement avec quelques exemples de paires question/réponse à partir du Web.

5.5. Les modèles

Une approche basée sur un modèle utilise des modèles préformâtes pour les questions. Cette approche vise beaucoup plus l'illustration que l'interprétation des questions et des réponses. L'ensemble des modèles est construit de façon à contenir le nombre optimal des modèles en s'assurant qu'il couvre adéquatement l'espace du problème. Le principe des SQR basés sur les modèles est très similaire au système de réponse automatisé FAQ (Foire aux questions) qui répond avec des réponses préenregistrées à la question de l'utilisateur, contrairement aux FAQ statiques.

6. Evaluation

Les systèmes QR répondent mieux à la demande et aux besoins des utilisateurs. Ils sont plus difficiles à évaluer que les autres types de systèmes informatiques, notamment les

moteurs de recherche. On s'accorde en effet sur les mesures d'évaluation des systèmes de recherche d'information, principalement la précision et le rappel. Évaluer les systèmes QR est une problématique, car les mesures dont nous disposons ne sont pas assez fines et évaluent plutôt une quantité, alors qu'une approche qualitative serait nécessaire, et il est difficile de définir ce qu'est une bonne réponse.

D'autre part, le manque de bases de test standardisées multi-langues et multi-domaines, notamment pour la langue arabe, reste un obstacle pour aboutir à une évaluation crédible et équitable.

Dans le reste de cette section, on présente quelques conférences d'évaluation. Prenant en exemple TREC, CLEF et QUALD.

6.1. TREC

Text Retrieval Conference (TREC), co-sponsorisée par National Institute of Standards and Technology (NIST) et le Département de la Défense des États-Unis, a été lancée en 1992 dans le cadre du programme TIPSTER Texte. Son but était de soutenir la recherche au sein de la communauté de recherche d'information en fournissant l'infrastructure nécessaire pour l'évaluation à grande échelle des méthodes liées à la recherche d'information :

En particulier, la série d'ateliers TREC a les objectifs suivants:

- encourager la recherche en recherche d'information en se basant sur de grandes collections de tests.
- accroître la communication entre l'industrie, le milieu universitaire et le gouvernement en créant un forum ouvert pour l'échange d'idées de recherche.
- accélérer le transfert de la technologie des laboratoires de recherche vers les produits commerciaux en démontrant des améliorations substantielles dans les méthodes de résolution des problèmes du monde réel.
- accroître la disponibilité de techniques d'évaluation appropriées pour l'industrie et les universités, y compris le développement de nouvelles techniques d'évaluation plus applicables aux systèmes actuels.

Un atelier TREC se compose d'un ensemble de sous-catégories, appelées tâches « tracks » dans la terminologie TREC, regroupées selon le domaine dans lequel ces tâches de recherche particulières sont définies. Dans la version 2017 de TREC, on distingue les tâches suivantes:

Noyau « Core » commun « Common Core » (trec-core 2017) : Il s'agit d'une nouvelle tâche pour TREC 2017. La piste servira de tâche commune à un large domaine de d'application en RI (recherche d'information), attirant ainsi un ensemble de ressources divers qui peuvent être utilisées pour étudier de nouvelles méthodologies pour la construction des collections de tests.

Recherche de réponse complexe (Dietz and Gamari 2017) : Il s'agit aussi d'une nouvelle tâche pour TREC 2017. Son objectif est de développer des systèmes capables de répondre à des questions complexes en rassemblant des informations pertinentes provenant d'un corpus entier.

Domaine dynamique (Grace Hui et al. 2017) : Cette tâche se focalise sur les algorithmes de recherche interactifs qui s'adaptent aux besoins d'information dynamique des utilisateurs professionnels, lorsqu'ils explorent des domaines complexes.

Live QA (Eugene et al. 2017) **:** Dans cette tâche, les systèmes génèrent des réponses à de vraies questions provenant d'utilisateurs réels via un flux de questions en temps réel.

OpenSearch Track (Schuth and Balog 2017) : La tâche OpenSearch explore le paradigme d'évaluation «Living Labs» pour l'RI qui implique de véritables utilisateurs de moteurs de recherche opérationnels.

Médecine de précision (Kirk et al. 2017) **:** Il se concentre sur la construction des systèmes qui utilisent des données (par exemple, les antécédents médicaux et les informations génomiques d'un patient) pour lier les patients en oncologie à des essais cliniques pour de nouveaux traitements, ainsi qu'une littérature factuelle pour identifier les traitements existants les plus efficaces.

Résumé en temps réel (Trec 2017) : La tâche RTS (Real-Time Summarization) explore les techniques de construction de résumés en temps réel à partir de flux de médias sociaux en réponse aux besoins d'information des utilisateurs.

Suivi des tâches (Emine et al. 2017) : Le suivi des tâches se situe à l'intersection de la diversité des sessions de TREC. Les principaux objectifs de la tâche consistent à tester si les systèmes peuvent induire les tâches possibles que les utilisateurs pourraient tenter d'accomplir en réponse à une requête.

Pour chacune, le **NIST** fournit une collection de documents, qui est utilisée par tous les participants de chaque piste « track » particulière. Peu de temps avant la conférence, les participants reçoivent un ensemble de ressources d'information, appelées thèmes « topics » dans la terminologie TREC.

6.1.1. SQR dans TREC

La tâche QR « QA track » est apparue en 1999 dans le cadre de la huitième conférence d’évaluation en recherche d’information, la TREC-8. Les candidats (entreprises ou laboratoires de recherche pour la plupart) doivent proposer un SQR capable de répondre à des questions de culture générale, triviales en apparence, du type :

How far is it from Denver to Aspen?

What county is Modesto, California in?

Who was Galileo?

What is an atom?

When did Hawaii become a state?

How tall is the Sears Building?

Listing 2. Exemple de questions triviales

Pour ce faire, ils ont accès, lors de la compétition, à une base composée d'une collection de documents issus de journaux de langue anglaise. Lors des campagnes d'évaluation de 1999 et 2000, il était demandé aux candidats de fournir une réponse de 250 ou 50 caractères (Voorhees Ellen M and Tice DM 2000).

Par exemple, pour la question *What are the animals that don't have backbones called?*

On obtient en exécution des réponses sous forme de phrases de 50 caractères (ou moins) du type: *backbones -- collectively called invertebrates -- ; Invertebrates*; *invertebrates-seem to have the kind of immune sys*.

Durant l'évolution de TREC, la tâche des SQR a changé de forme et d'objectif. Actuellement, la version 2017 de QA TREC propose Live QA, la nouvelle tâche d'évaluation des SQR introduite en 2015, comme l'évolution de la tâche traditionnelle question-réponse.

La tâche Live QA se reproduira en 2017, en se concentrant sur les questions en temps réel répondant aux questions de l'utilisateur provenant du flux en direct des questions les plus récentes soumises au site Yahoo Answers, ainsi qu'une nouvelle sous-tâche sur les questions médicales.

Dans le Figure 4, des extraits de bases de tests pour les compétions QA 2007 et Live QA 2016 :

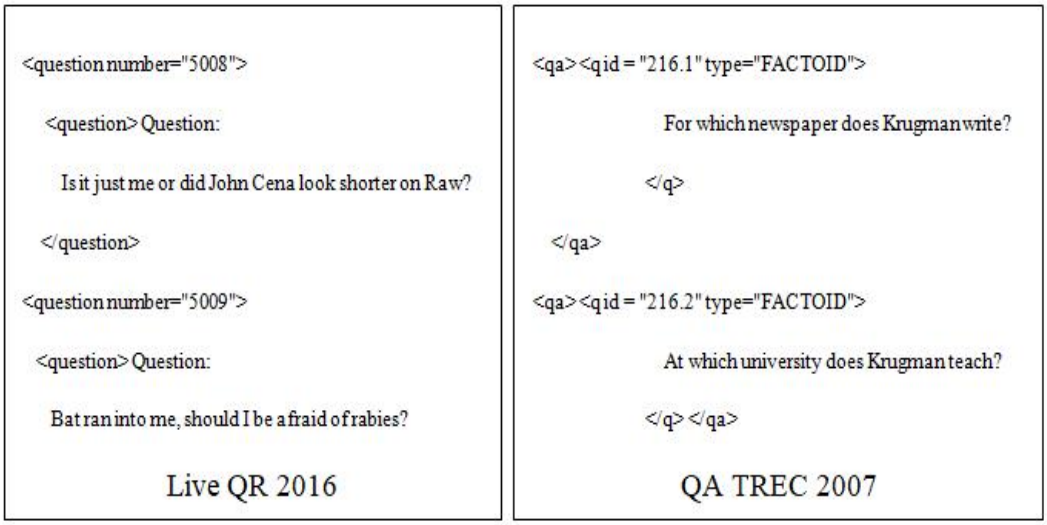

Figure 4: Extrait de test TREC

6.2. CLEF

L'initiative CLEF (Conference and Labs of the Evaluation Forum, connu aussi ***Cross - Language Evaluation Forum***), est un organisme autogéré dont la mission principale est de promouvoir la recherche, l'innovation et le développement de systèmes d'accès à l'information en mettant l'accent sur le multilinguisme et l'information multimodal avec différents niveaux de structure. CLEF favorise la recherche et le développement en fournissant une infrastructure pour (PROMISE 2017) :

- Tester, exécuter et évaluer les systèmes multilingues et multimodaux.
- Étude de l'accès à l'information par l'utilisation des données non-structurées, semi-structurées, hautement structurées et sémantiquement enrichies.
- Création de collections de tests réutilisables pour l'analyse comparative.
- L'exploration de nouvelles méthodes d'évaluation et de moyens novateurs pour utiliser des données expérimentales.
- Discussion des résultats, comparaison des approches, échange d'idées et transfert de connaissances.

6.3. Les QALD

Les dernières années ont vu un nombre croissant de recherches sur les SQR (systèmes question-réponse) sur les données du Web sémantique, formant un paradigme d'interaction qui permet aux utilisateurs finaux de tirer parti de la puissance des standards du Web sémantique tout en dissimulant leur complexité et interface facile à utiliser. Parallèlement, la quantité croissante de données a conduit à un paysage de données hétérogène où les SQR ont du mal à suivre le volume, la variété et la véracité des connaissances sur les données liées.

Le challenge QALD (Question Answering over Linked Data) vise à fournir une référence à jour pour évaluer et comparer les systèmes de pointe qui servent d'intermédiaire entre un utilisateur, exprimant son besoin d'information en langage naturel et les données RDF. Il s'adresse donc à tous les chercheurs et praticiens travaillant sur l'interrogation des données

liées en utilisant : le traitement du langage naturel pour répondre aux questions, la récupération multilingue d'information et les sujets connexes. L'objectif principal est d'obtenir des informations sur les forces et les faiblesses des différentes approches et sur les solutions possibles pour faire face à la nature large, hétérogène et distribuée des données du Web sémantique.

Le principal défi du QR sur les données liées est de traduire le besoin d'information d'un utilisateur (requête de l'utilisateur) en une forme évaluable à l'aide des techniques standards de traitement des requêtes et d'inférence du Web sémantique. La tâche principale de QALD est donc la suivante :

> *Avec un ou plusieurs jeux de données RDF ainsi que des sources de connaissances supplémentaires et des questions ou mots-clés en langage naturel, renvoyer les réponses correctes ou une requête SPARQL qui récupère ces réponses.*

Pour la version 2017 du challenge QUALD on distingue trois tâches :

Tâche 1 : question-réponses multilingues à DBpedia

Compte tenu de la diversité des langues utilisées sur le Web, il est nécessaire de faciliter l'accès multilingue aux données sémantiques. La tâche principale de QALD consiste donc à extraire les réponses d'un référentiel de données RDF en fonction d'un besoin d'information exprimé en une variété de langues naturelles.

Les données de formation comporteront plus de 500 questions compilées et élaborées à partir des défis précédents. Les questions seront disponibles en huit langues différentes (anglais, espagnol, allemand, italien, français, néerlandais, roumain et farsi), avec éventuellement trois langues supplémentaires (coréen, hindi et portugais brésilien). Ces questions sont des questions factuelles générales à domaine ouvert, par exemple :

(*en) Which book has the most pages?*

(de) Welches Buch hat die meisten Seiten?

(es) Que libro tiene el mayor numero de paginas?

(it) Quale libro ha il maggior numero di pagine?

(fr) Quel livre a le plus de pages?

(nl) Welk boek heeft de meeste pagina's?

(ro) Ce carte are cele mai multe pagini?

Listing 3. Exemple de questions factuelles en différentes langues

Les questions varient en fonction de leur complexité, y compris des questions avec des chiffres (par exemple, combien Mohamed a-t-il d'enfants?), des superlatifs (par exemple, quel musée à New York compte le plus de visiteurs?), et agrégateurs temporels (par exemple, combien de sociétés ont été

créées la même année que Google?). Chaque question est annotée par une requête SPARQL spécifiée manuellement et par la réponse correspondante.

Tâche 2 : Question-réponse hybride

De nombreuses informations sont toujours disponibles uniquement sous forme de texte, à la fois sur le Web, et sous forme d'étiquettes et de résumés dans les sources de données liées. Par conséquent, il faut des approches qui peuvent non seulement traiter le caractère spécifique des données structurées, mais également trouver des informations dans plusieurs sources, traiter des informations structurées et non-structurées et combiner ces informations rassemblées en une seule réponse.

QALD inclut donc une tâche sur la réponse hybride aux questions, demandant aux systèmes de récupérer les réponses aux questions nécessitant l'intégration des données à la fois de RDF et de sources textuelles.

Tâche 3 : Question-réponse à grande échelle répondant à RDF

L'accent sera mis sur l'inclusion d'un ensemble de questions à grande échelle. Les approches réussies peuvent évoluer vers un volume de données volumineux, traiter un grand nombre de questions et accélérer le processus de réponse aux questions par une parallélisassions, de sorte que le plus grand nombre de questions possibles puisse être répondu aussi précisément que possible dans les plus brefs délais.

Tâche 4: Question-réponse anglaise à Wikidata

Une autre nouvelle tâche introduite en 2017 utilise une source de données publique Wikidata en tant que référentiel cible. Les données de formation comprendront 100 questions factuelles de domaine ouvert compilées à partir de l'itération précédente de la tâche 1. Dans cette tâche, vous devez répondre aux questions formulées à l'origine pour DBpedia à l'aide de Wikidata. Ainsi, vos systèmes devront gérer une structure de représentation des données différente. Cette tâche aidera à évaluer le caractère générique de votre approche et sa facilité d'adaptation à une nouvelle source de données. Notez que les résultats obtenus à partir de Wikidata peuvent être différents des réponses aux mêmes requêtes trouvées dans DBpedia.

7. Conclusion

Dans ce chapitre, nous avons donné une vue globale sur les systèmes question-réponse. Nous avons retracé l'historique de l'évolution qui remonte aux années soixante. Dans (Simmons 1965) on trouve le premier état de l'art sur les systèmes question-réponse publié en 1965. Nous avons décrit l'architecture globale des systèmes question-réponse composé de plusieurs modules sur plusieurs sources de données, et nous avons fourni une classification de ces systèmes en se basant sur plusieurs critères. Nous avons discuté diverses approches pour les systèmes question-réponse. Enfin, nous avons discuté l'évaluation de ces systèmes en mettant l'accent sur les compagnes d'évaluation les plus importantes : TREC (Text Retrieval Conference) la version 2017, CLEF (Conference and Labs of the Evaluation Forum, connue aussi Cross -Language Evaluation Forum) la version 2016 et QUALD (Question Answering over Linked Data) la version 2017.

Chapitre 3 : Traitement automatique du langage naturel et Web sémantique

1. Introduction

Les bases de connaissances du Web sémantique sont généralement représentées sous forme des annotations sémantiques relatives à des ontologies. Leur interrogation passe par un langage de requête SPARQL, langage non maîtrisé par les utilisateurs non experts, qui requièrent de connaître le schéma de la base. C'est pourquoi les systèmes d'interrogation en langage naturel se développent actuellement. Ainsi, il se pose le problème de construction automatique des requêtes, prenant en considération les distances lexicales entre les mots de la question au vu de l'ontologie du domaine. D'autres problèmes considérables sont rencontrés lorsqu'on traite une langue particulière comme l'Arabe.

Pour les systèmes question-réponse Arabes dans le Web sémantique, on distingue deux grandes problématiques de recherche, à savoir le traitement automatique de la langue (TAL) arabe et les données liées arabes. La compatibilité et l'évolution de la recherche de ces deux technologies pour la langue arabe soufre d'une grande insuffisance.

Dans ce chapitre, nous présentons un background pour le développement d'un système question-réponse arabe pour le Web sémantique. La section suivante présente le Web sémantique. La section 3 donné un aperçu sur la langue et le TAL arabe. La section 4 introduit la recherche sur le Web sémantique arabe avec un exemple du monde réel qui est le projet DBpedia et son chapitre arabe.

2. Le Web sémantique

2.1. Le Web sémantique extension du Web actuel

En 1989, Tim Berners-Lee, futur fondateur du W3C, imagine pour le CERN, une organisation novatrice des ressources internes sous forme d'un réseau informatique distribué. La structure proposée est celle d'un graphe dont les nœuds représentent les ressources (personnes, groupes de personnes, projets, concepts, documents, objets du hardware, entités de natures diverses), et les arcs représentent des liens étiquetés connectant les nœuds. Des exemples d'étiquettes proposées, en 1989, sont : *dépend de*, *est partie de*, *réalise*, *réfère à*, *utilise*, *est un exemple de*, etc. Cette vision, reprenant la notion d'hypertexte est à l'origine du Web. Pourtant, malgré la rapidité de son expansion avec un nombre impressionnant de sites et d'utilisateurs, le Web actuel ne reflète qu'imparfaitement la vision initiale. Le Web ne connecte que des documents (nœuds non typés), et les liens eux-mêmes ne sont pas munis d'étiquettes typées. Les utilisateurs humains ont appris à faire avec ces limitations en s'appuyant sur leur compréhension du contenu des documents connectés et en attribuant, si possible, un sens aux liens à l'aide du texte des ancres. Par contre, ces capacités restent largement inaccessibles aux logiciels, faute d'une sémantique interprétable par les machines. Il est certain que cette volumineuse donnée augmente constamment la difficulté de trouver, d'accéder et de présenter les informations requises par un large éventail d'utilisateurs. C'est principalement à cause de la représentation de l'information sous une forme lisible par l'homme, plutôt que d'être lisible par machine aussi. Cela signifie que, parmi un tas d'informations pouvant être extrait par la machine, l'utilisateur lui-même doit effectuer la prochaine étape de filtrage. Il doit extraire ses propres informations désirées ou ce qu'il

cherche exactement dans ce tas. En fait, le Web ne peut être pleinement fonctionnel que lorsque ses données peuvent être trouvées, partagées, traitées et comprises par l'homme et par la machine (Davies et al. 2003). Cette vision du Web peut être atteinte par la nouvelle génération du Web, c'est-à-dire Web 3.0 ou WS, qui est la solution prometteuse pour réduire cet écart entre la forme d'information lisible par l'homme et cette machine compréhensible. Il représente le contenu Web sous une forme plus facile à être traitée par la machine et utilise des techniques intelligentes pour tirer parti de cette représentation. C'est en ajoutant une couche sémantique aux pages Web existantes pour décrire leur contenu, de sorte que la page Web contiendra les informations de formatage pour la présentation au lecteur humain, ainsi que les informations sur leur contenu (les métadonnées) pour la compréhension de la machine. Puisque les «métadonnées» comprennent un fragment de la signification des données, le terme «sémantique» est apparu et «WS» s'est matérialisé. L'approche utilisée pour ajouter cette couche sémantique est l'annotation sémantique (Horrocks 2008), qui consiste à étiqueter les pages Web avec la sémantique de leur contenu en mappant les instances de données à des concepts ontologiques dans une ontologie prédéfinie.

2.2. Définition

Citant Tim Berners-Lee, l'inventeur du World Wide Web : «Pour un ordinateur, le Web est un monde plat et ennuyeux sans signification. Une nouvelle forme de contenu Web significatif pour les ordinateurs déclenchera une révolution de nouvelles possibilités. Le Web sémantique (WS) n'est pas un Web séparé, mais une extension de l'actuel, dans lequel l'information est définie de façon précise. L'ajout de la sémantique au Web implique l'autorisation de documents contenant des informations dans des formulaires lisibles par une machine, et permettant de créer des liens avec des valeurs de relation».

"The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

-- Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, Mai 2001.

2.3. Le modèle en couche du Web sémantique

La Figure 5 montre l'approche en couches du Web sémantique qui décrit la vision du Web sémantique. Pour rendre l'approche Web sémantique plus acceptable par les organisations, les entreprises, les groupes et les utilisateurs individuels, elle doit diviser la vision Web sémantique en plus petites parties, étapes ou couches. Par conséquent, le modèle Web sémantique est une collection de couches qui sont sous une forme hiérarchique, c'est-à-dire, chacune est une couche au-dessus d'une autre couche.

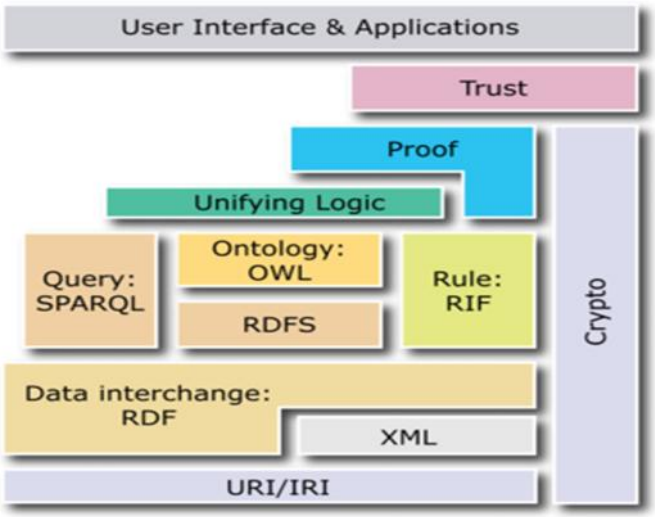

Figure 5: Le modèle en couche du Web sémantique

Plusieurs formats et langages forment les couches du Web sémantique. Nous trouvons des Identificateurs exprimés par des URI (Uniform Resource Identifier), des Documents écrits en XML (Extensible Markup Language), des métadonnées écrites en RDF (Resource Description Framework). Nous trouvons aussi divers formats d'échange de données (par exemple RDF/XML, N3), des notations, telles que RDF Schemas (RDFS) et le langage d'ontologie Web (OWL), qui sont tous destinées à fournir une description formelle des concepts, des termes et des relations dans un domaine de connaissances donné. Dans ce qui suit, nous détaillons les éléments des couches du WS :

URI :

Pour identifier les éléments dans le Web, nous utilisons des identificateurs, parce que nous utilisons un système d'identification uniforme, et parce que chaque élément identifié est considéré comme une «ressource». Nous appelons ces identificateurs «Uniform Resource Identifiers» ou URI en abrégé. Nous pouvons donner un URI à n'importe quoi, et tout ce qui a une URI peut être dit "dans le Web" : une personne, un livre que vous avez acheté la semaine dernière, et tout ce que vous pouvez penser - - ils peuvent tous avoir une URI.

XML :

XML a été conçu pour être un moyen simple d'envoyer des documents sur le Web. Il permet à n'importe qui de concevoir son propre format de document, puis d'écrire un document dans ce format. Ces formats de document peuvent inclure un balisage pour améliorer la signification du contenu du document. Ce balisage est "lisible par machine", c'est-à-dire que les programmes peuvent le lire et le comprendre. En incluant une signification lisible par machine dans nos documents, nous les rendons beaucoup plus puissants.

RDF :

Le bloc de construction le plus fondamental est le Framework de Description de Ressource (RDF), un format pour définir des informations sur le Web. RDF est un langage de balisage pour décrire les informations et les ressources sur le Web. La mise à disposition d'information

dans des fichiers RDF permet aux programmes informatiques («Web spiders») de rechercher, découvrir, récupérer, collecter, analyser et traiter des informations sur le Web. Le Web sémantique utilise RDF pour décrire les ressources Web. RDF fournit un modèle pour les données, et une syntaxe permettant aux parties indépendantes de les échanger et de les utiliser. Il est conçu pour être lu et compris par les ordinateurs. Il n'est pas conçu pour être exposé aux humains.

RDF est vraiment très simple. Une déclaration RDF ressemble beaucoup à une phrase simple, sauf que presque tous les mots sont des URI. Chaque déclaration RDF comporte trois parties : sujet, prédicat et objet. Regardons une simple déclaration RDF:

```
<http://aaron.com/>

<http://love.example.org/terms/reallyLikes>

<http://www.w3.org/People/Berners-Lee/Weaving/>.
```

Listing 4. Une simple déclaration RDF

Le premier URI est le sujet. Dans ce cas, le sujet est Aaron. Le deuxième URI est le prédicat. Il relie le sujet à l'objet. Dans ce cas, le prédicat est "aime vraiment". Le troisième URI est l'objet. Ici, l'objet est le livre de Tim Berners-Lee "Weaving". Donc, la déclaration RDF ci-dessus dit que : Aaron aime vraiment "Weaving".

RDFS :

Le schéma RDF a été conçu pour être un modèle de type de données simple pour RDF. En utilisant RDF Schema, nous pouvons dire que "Fido" est un type de "Chien", et que "Chien" est une sous-classe de ″Animal″. Nous pouvons également créer des propriétés et des classes, ainsi que faire un peu plus de choses avancées, telles que la création des Ranges et des Domaines pour les propriétés.

Les trois premiers concepts les plus importants que RDF et RDF Schema nous donnent sont la "Ressource" (rdfs:Resource), la "Classe" (rdfs:Class) et la "Propriété" (rdf:Property). Ce sont toutes des classes. Par exemple, tous les termes de RDF sont des types de ressources. Pour déclarer que quelque chose est un type d'autre chose, nous utilisons simplement la propriété ″rdf:type″.

```
rdfs:Resource rdf:type rdfs:Class .
rdfs:Class rdf:type rdfs:Class .
rdf:Property rdf:type rdfs:Class .
rdf:type rdf:type rdf:Property .
```

Listing 5. Description de quelques concepts RDF et RDFS

Listing .5 indique simplement que "resource est un type de classe, la classe est un type de classe, la propriété est un type de classe et le type est un type de propriété". Ce sont toutes les vraies déclarations.

Ontologies :

L'ontologie est l'une des pierres angulaires des applications WS. Pour un certain domaine, l'ontologie énumère et donne des descriptions sémantiques des concepts dans ce domaine, définissant les attributs de concepts pertinents pour le domaine, et diverses relations entre eux. L'ontologie est un document ou un fichier qui définit formellement les relations entre les termes. Une ontologie est une description explicite d'un domaine. Elle comprend :

- les concepts,
- les propriétés et les attributs des concepts,
- les contraintes sur les propriétés et les attributs
- et les individus (souvent, mais pas toujours).

Une ontologie définit un vocabulaire commun, i.e., une compréhension partagée. Comme on le sait, en général « l'ontologie est une spécification explicite d'une conceptualisation » (Gruber 1993). L'ontologie est aussi définie dans la communauté informatique comme « une compréhension partagée et commune d'un domaine qui peut être utilisée par des personnes et des systèmes hétérogènes et distribués » (Fensel 2000).

Le W3C a défini l'ontologie comme suit : « Une ontologie définit les termes utilisés pour décrire et représenter un domaine de connaissance. Les ontologies sont utilisées par des personnes, des bases de données et des applications qui ont besoin de partager des informations de domaine (un domaine spécifique ou un domaine de connaissances, comme la médecine, la fabrication d'outils, l'immobilier, la réparation automobile, la gestion financière, etc. » (Heflin 2004).

SPARQL :

SPARQL est un langage de requête RDF, et c'est l'un des composants centraux du Web sémantique. SPARQL signifie SPARQL Protocol et RDF Query Language. Sa définition simple est la suivante : « SPARQL est un langage de requête que nous pouvons utiliser pour interroger le contenu des données RDF. SPARQL fournit également un protocole que nous devons suivre si nous voulons interroger un ensemble de données RDF» (Yu 2011).

Le W3C a défini le SPARQL comme suit : « SPARQL peut être utilisé pour exprimer des requêtes sur diverses sources de données. Ces données sont soient stockées nativement en tant que RDF ou vues en tant que RDF via un middleware. SPARQL prend également en charge l'agrégation, les sous-requêtes, la création de valeurs par des expressions et le graphe RDF source. Les résultats des requêtes SPARQL peuvent être des ensembles de résultats ou des graphes RDF » (Harris et al. 2013)

Les avantages de SPARQL sont résumés comme suit :

- Obtenir des informations particulières via des graphes de requête RDF.
- Génération de rapports en exécutant à nouveau des requêtes RDF.
- Permettre aux applications de traiter directement les résultats des requêtes SPARQL au lieu des documents RDF.

Règles d'inférence :

Les règles d'inférence permettent d'inférer des conclusions basées sur des règles et des faits disponibles dans la base de connaissances. Cela augmente donc encore plus la puissance du Web sémantique. Par exemple, une ontologie peut exprimer la règle "Si un code de ville est associé à un code d'état, et qu'une adresse utilise ce code de ville, alors cette adresse a le code d'état associé". Un programme pourrait alors facilement déduire, par exemple, qu'une adresse de Cornell University, située à Ithaca, doit se trouver dans l'État de New York, qui se trouve aux États-Unis, et devrait donc être conforme aux normes américaines. L'ordinateur ne "comprend" pas vraiment cette information, mais il peut maintenant manipuler les termes beaucoup plus efficacement de manière utile et significative pour l'utilisateur humain.

Preuve :

Une fois que nous commençons à construire des systèmes qui suivent la logique, il est logique de les utiliser pour prouver des choses. Les gens du monde entier pouvaient écrire des déclarations logiques. Ensuite, votre machine pourrait suivre ces "liens" sémantiques pour construire des preuves.

Par exemple, les enregistrements de ventes de sociétés montrent que Jane a vendu 55 widgets et 66 pignons. Le système d'inventaire indique que les widgets et les pignons sont deux produits de société différents. Les règles mathématiques intégrées stipulent que 55 + 66 = 121 et que 121 est supérieur à 100. Et, comme nous le savons tous, une personne qui vend plus de 100 produits est membre du club Super Salesman. L'ordinateur réunit toutes ces règles logiques pour prouver que Jane est un super vendeur.

Bien qu'il soit très difficile de créer ces preuves (il peut être nécessaire de suivre des milliers, voire des millions de liens dans le Web sémantique), cela n'est généralement pas nécessaire car les informations sur le Web n'ont pas à être prouvées.

Confiance (signatures numériques et Web de confiance) :

Maintenant, nous pouvons dire que tout ce plan est génial. Qui aurait confiance en un tel système? C'est là que la signature numérique entre en jeu.

Sur la base des travaux en mathématiques et en cryptographie, les signatures numériques prouvent qu'une personne a écrit (ou est d'accord avec) un document ou une déclaration. Donc, on signe numériquement toutes leurs déclarations RDF. De cette façon, nous pouvons être sûrs qu'il les a écrits (ou du moins, ils ont garanti leur authenticité). Maintenant, nous pouvons simplement dire à notre programme quelles signatures doivent faire confiance.

3. La langue arabe

3.1.La langue arabe une langue sémitique

L'arabe est l'une des langues sémitiques qui sont l'hébreu, l'acadien, le phénicien, le tigre, l'araméen, le syriaque, ougaritique, l'amharique, le geez et le tigina. Toutes ces langues sémitiques sont mortes, ou du moins ne sont utilisées que de manière simple, sauf l'arabe. Par exemple, les langues acadiennes et ougaritiques sont mortes il y a longtemps. L'hébreu est l'une des plus anciennes langues sémitiques qui a disparu, mais qui s'est récemment

développée en Palestine par l'entité sioniste. La langue tigre est une langue utilisée uniquement comme langue liturgique de l'église éthiopienne et érythréenne orthodoxe de Tewahedo. Geez, qui était la langue officielle du royaume d'Aksoum et de la cour impériale éthiopienne, n'est utilisée que dans la littérature de l'Église éthiopienne orthodoxe de Tewahedo et de l'Église éthiopienne.

Il y a beaucoup de raisons pour lesquelles l'arabe est toujours utilisée et a beaucoup de locuteurs jusqu'à présent. La première raison est la prospérité de la langue arabe à l'ère Jāhiliyyah (barbare, primitive) de la société arabe à cause de l'art de la poésie. Il y avait beaucoup de poèmes écrits par des poètes arabes. Chaque tribu encourageait ses membres à apprendre l'arabe pour devenir des poètes habiles. Par conséquent, l'ère Jāhiliyyah était l'époque la plus riche des locuteurs de l'arabe et des poètes. A cette époque, certains poètes composaient le Mu'allaqāt, un poème contenant au moins mille vers. Il y avait aussi des concours annuels pour choisir le meilleur poème qui avait le plus d'éloquence et de sens. Ces concours ont contribué à la survie de la langue arabe. La deuxième raison était la mission du prophète Mohamed (paix et bénédictions soient sur lui) et le Coran. Le Coran est le livre saint pour les musulmans, qui est en langue arabe. Le prophète Mohamed (paix et bénédictions soient sur lui) enseignait les règles islamiques en arabe et la plupart des rituels fondamentaux liés au culte en islam, qui devaient être faits en arabe pour les peuples arabes ou non arabes. Par exemple, il y a cinq prières quotidiennes dans lesquelles certaines parties du Coran doivent être lues en arabe. Ce sont les deux principales raisons pour lesquelles la langue arabe a survécu et a incité les non-Arabes à apprendre les bases de la langue arabe (Salahudeen et al. 2010) .

3.2. Le système d'écriture

L'arabe s'écrit de droite vers la gauche en utilisant deux types de caractères : les lettres et les diacritiques. Les lettres arabes renferment les consonnes et les voyelles longues alors que les diacritiques servent à vocaliser le texte, i.e., déterminer la phonétique de ses mots. Quelques diacritiques, à savoir فتحة\fat.Haħ\, ضمة\Dam~aħ\ et كسرة\kas.raħ\ sont considérés comme des voyelles courtes. Dans ce système d'écriture, les lettres sont obligatoires alors que les diacritiques sont optionnelles. L'omission des diacritiques dans le texte arabe augmente le degré d'ambiguïté pour les analyseurs morphologiques et syntaxiques.

Dans la littérature du TAL, il y a un désaccord sur le nombre de lettres arabes. Ce désaccord est dû à la classification de ce qui est ou non diacritique et à l'ignorance de certaines lettres.

Généralement on parle de 28 lettres mais les nombres 36 et 40 sont considérés aussi dans la littérature du domaine (Buckwalter 2004b; Habash 2010). Les 36 lettres de l'ASM sont classées comme suit :

• 28 lettres de base,

• 6 formes possibles de la lettre Hamza,

• la lettre Ta-Marbouta

• et la lettre Alif-Maqsoura.

En plus de ce nombre de lettres, on trouve dans les textes d'ASM d'autres lettres non arabes, telles que چ ,پ , گ pour représenter une phonétique non arabe correspondant aux lettres latines (p, j, v, g) qui n'existent pas originellement en ASM (Buckwalter 2004b). Ces lettres sont empruntées aux alphabets des langues écrites en alphabet arabe, telles que le persan/farsi.

En arabe, les diacritiques sont des symboles optionnels dans l'écriture, i.e. que l'écriture arabe peut se trouver complètement, partiellement ou non diacritisée. L'ajout de diacritiques dans un texte a pour but d'aider le lecteur à lire et prononcer correctement les mots (Doumi 2017).

3.3. Eléments du script arabe

Il existe deux types de symboles arabes pour écrire des mots : les lettres et les signes diacritiques. En plus de ces symboles, nous présentons dans cette section des chiffres, de la ponctuation et d'autres symboles.

3.3.1. Lettres

Les lettres arabes sont écrites en style cursif à la fois en caractères d'imprimerie et en écriture (écriture manuscrite). Elles se composent généralement de deux parties : la forme de lettre (رسم rasm) et la marque de lettre (إعجام AiçjAm). La forme de lettre est un composant essentiel dans chaque lettre. Il y a un total de 19 formes de lettres (voir la Figure 6). Les marques de lettre, également appelées signes diacritiques consonnes, peuvent être subdivisées en trois types (voir la Figure 7). Il y a d'abord les points, avec cinq possibilités : un, deux ou trois pour aller au-dessus de la forme de lettre et un ou deux pour passer sous la forme de lettre. Deuxièmement, le court Kaf, utilisé pour marquer des formes de lettres spécifiques de la lettre Kaf (voir le Tableau 1). La troisième est la lettre Hamza (همزة Hamza¯ h). Le Hamza peut apparaître au-dessus ou au-dessous des formes de lettres spécifiques. Le terme Hamza est utilisé à la fois pour la forme de lettre (ء) et la lettre, qui apparaît avec d'autres formes de lettres, telles que أ Â, ؤ W et? ئ Y. La lettre de Madda (مدةMad~a¯ h) est une variante de Hamza.

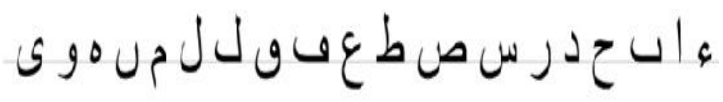

Figure 6: Formes de lettres dans la langue arabe

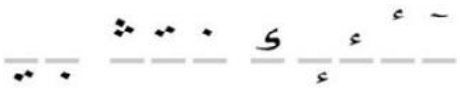

Figure 7: Les marques de lettre pour la désambigüisation de différentes lettres

3.3.2. Diacritiques

La deuxième classe de symboles dans l'écriture arabe est les signes diacritiques. Alors que les lettres sont toujours écrites, les signes diacritiques sont facultatifs. L'arabe écrit peut être complètement diacritique, partiellement diacritique ou entièrement non-diacritique. Le texte arabe n'est pas critiqué, sauf dans les textes religieux, les textes éducatifs pour enfants et certains poèmes. Certains diacritiques sont indiqués en arabe écrit moderne pour aider les lecteurs à désambiguïser certains mots. Dans le Penn Arabic Treebank (Maamouri et al. 2004), 1,6% de tous les mots ont au moins un signe diacritique indiqué par leur auteur. 99,3% d'entre eux sont réellement corrects, car ils apparaissent dans la position correcte dans le mot.

Il existe trois types de signes diacritiques : la voyelle, la nonne et la Shadda. Ils sont présentés dans la Figure 9. Les signes diacritiques des voyelles représentent les trois voyelles courtes de l'arabe (Fatha / ◌َ /, Damma /◌ُ / et Kasra / ◌ِ /) et l'absence de voyelle (pas de voyelle, Sukun ◌ْ).

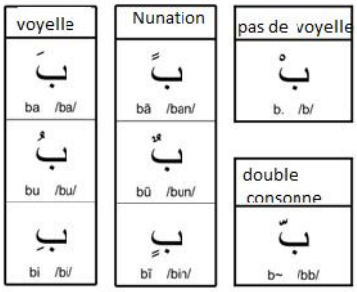

Figure 8: Types de diacritiques arabes

3.3.3. Les chiffres

Les chiffres arabes sont écrits dans un système décimal. Il existe deux séries de chiffres pour écrire des nombres dans le monde arabe. Les chiffres arabes couramment utilisés en Europe, en Amérique et dans les pays arabes occidentaux (Maroc, Algérie, Tunisie). Les pays arabes du Moyen-Orient (par exemple, l'Égypte, la Syrie, l'Iraq et l'Arabie saoudite) utilisent ce que l'on appelle les chiffres indo-arabes. Les pays non arabes, tels que l'Iran et le Pakistan, utilisent une variante du jeu de chiffres indo-arabes, qui ne diffèrent que par les chiffres 4, 5 et 6. Les ensembles à trois chiffres sont contrastés dans la Figure 10.

Arabe maghreb algerie-maroc tunis..	0	1	2	3	4	5	6	7	8	9
Indo arabe moyen-orient	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
Indo-arabe oriental Iran, pakistan..	٠	١	٢	٣	۴	۵	۶	٧	٨	٩

Figure 9: Trois types de chiffres utilisés dans le script arabe

Bien que l'arabe soit écrit de droite à gauche, les formes de nombres à plusieurs chiffres en arabe sont les mêmes que celles utilisées dans les langues européennes (de gauche à droite).

En tapant, les nombres à plusieurs chiffres sont saisis de gauche à droite (voir la Figure 11). En écriture manuscrite, un nombre à deux chiffres est écrit de droite à gauche, mais des nombres plus importants commencent à gauche et se dirigent vers la droite. Ceci reflète la manière dont les chiffres arabes sont couramment prononcés en arabe : en plus petits nombres (jusqu'à 100), le chiffre de plus petite valeur de lieu est prononcé (et écrit en premier), mais en plus grand nombre, la plus haute valeur de lieu est prononcée. Par exemple, un nombre tel que 2345 est prononcé : deux mille trois cent cinq et quarante. Le mappage entre chiffres et énoncés est important pour des applications, telles que la synthèse vocale, et également pour la modélisation du langage pour la reconnaissance automatique de la parole (ASR) (Habash and Roth 2008).

Figure 10: Les chiffres arabes sont affichés de gauche à droite dans un texte de droite à gauche

3.4. Codages arabes

De nombreux encodages "standard" ont été développés pour l'arabe au fil des années. Nous ne discutons ici que les trois encodages les plus couramment utilisés, qui sont tous bien pris en charge pour les entrées et les sorties sur différentes plates-formes. Le Tableau 1 présente les différentes valeurs de code utilisées pour les symboles arabes MSA dans plusieurs codages côte à côte. Pour des discussions supplémentaires sur les normes d'encodage en arabe, voir (Borras 2004; Engström 2008).

Unicode	Forme Contextuelle				Nom
	Isolé	Fin	Milieu	Début	
0627 ا	FE8D ا	FE8E ا			*ʾalif*
0628 ب	FE8F ب	FE90 ب	FE92 بـ	FE91 بـ	*bāʾ*
062A ت	FE95 ت	FE96 ت	FE98 تـ	FE97 تـ	*tāʾ*
062B ث	FE99 ث	FE9A ث	FE9C ثـ	FE9B ثـ	*ṯāʾ*
062C ج	FE9D ج	FE9E ج	FEA0 جـ	FE9F جـ	*ǧīm*
062D ح	FEA1 ح	FEA2 ح	FEA4 حـ	FEA3 حـ	*ḥāʾ*
062E خ	FEA5 خ	FEA6 خ	FEA8 خـ	FEA7 خـ	*ḫāʾ*
062F د	FEA9 د	FEAA د			*dāl*
0630 ذ	FEAB ذ	FEAC ذ			*ḏāl*
0631 ر	FEAD ر	FEAE ر			*rāʾ*
0632 ز	FEAF ز	FEB0 ز			*zayn/zāy*
0633 س	FEB1 س	FEB2 س	FEB4 سـ	FEB3 سـ	*sīn*
0634 ش	FEB5 ش	FEB6 ش	FEB8 شـ	FEB7 شـ	*šīn*
0635 ص	FEB9 ص	FEBA ص	FEBC صـ	FEBB صـ	*ṣād*
0636 ض	FEBD ض	FEBE ض	FEC0 ضـ	FEBF ضـ	*ḍād*
0637 ط	FEC1 ط	FEC2 ط	FEC4 طـ	FEC3 طـ	*ṭāʾ*
ظ	FEC5 ظ	FEC6 ظ	FEC8 ظـ	FEC7 ظـ	*ẓāʾ*
0639 ع	FEC9 ع	FECA ع	FECC عـ	FECB عـ	*ʿayn*
063A غ	FECD غ	FECE غ	FED0 غـ	FECF غـ	*ġayn*
0641 ف	FED1 ف	FED2 ف	FED4 فـ	FED3 فـ	*fāʾ*
0642 ق	FED5 ق	FED6 ق	FED8 قـ	FED7 قـ	*qāf*
0643 ك	FED9 ك	FEDA ك	FEDC كـ	FEDB كـ	*kāf*
0644 ل	FEDD ل	FEDE ل	FEE0 لـ	FEDF لـ	*lām*
0645 م	FEE1 م	FEE2 م	FEE4 مـ	FEE3 مـ	*mīm*
0646 ن	FEE5 ن	FEE6 ن	FEE8 نـ	FEE7 نـ	*nūn*
0647 ه	FEE9 ه	FEEA ه	FEEC هـ	FEEB هـ	*hāʾ*
0648 و	FEED و	FEEE و			*wāw*
064A ي	FEF1 ي	FEF2 ي	FEF4 يـ	FEF3 يـ	*yāʾ*
0622 آ	FE81 آ	FE82 آ			*ʾalif madda*
0629 ة	FE93 ة	FE94 ة	—	—	*Tāʾ marbūṭ*
0649 ى	FEEF ى	FEF0 ى	—	—	*ʾalif maqṣūr*

Tableau 1: Valeurs de code utilisées pour les symboles arabes MSA

3.4.1. Codages 8 bits: ISO-8859-6 et CP-1256

ISO-8859-6 et CP-1256 sont deux des schémas d'encodage précoce les plus populaires de l'arabe.

ISO-8859-6, développé par l'organisation internationale de normalisation (ISO), est identique à la norme ASMO-708 créée par les normes arabes et l'organisation de métrologie (ASMO). CP-1256 (Page de code 1256), l'alias l'encodage arabe de Windows, a été développé par Microsoft et a été rendu extrêmement populaire via Windows. Ces deux codages utilisent 1 octet (8 bits) pour représenter chaque symbole (pour un maximum de 256 caractères). Comme dans les autres encodages de leur classe pour les scripts/langages autres que l'arabe, les premiers 7 bits (ou 128 caractères) sont réservés à l'anglais ASCII (American Standard Code for Information Interchange). L'autre script est représenté dans les 128 autres caractères. Cela permet d'utiliser le même encodage pour deux scripts (ou plusieurs langues) si nécessaire. La partie arabe dans ASMO-708 et ISO-8859-6 est basée sur (compatible mais non identique à) un codage arabe antérieur de 7 bits (ASMO-449) (Engström 2008).

Dans CP-1256, les caractères arabes sont listés dans l'ordre, bien qu'il y ait des lacunes entre les différents ensembles de caractères afin de conserver les valeurs de code pour certaines langues européennes, en particulier le français, produisant ainsi une page de codes multilingue (arabe, anglais, français). Les CP-1256 et ISO-8859-6 ne pouvaient pas prendre en charge l'ensemble complet de caractères arabes étendus. Cependant, les caractères du Persan sont inclus. Ces codages spécifient uniquement les graphèmes et s'appuient sur des algorithmes distincts pour afficher les glyphes de police corrects.

CP-1256 et ISO-8859 ne sont pas compatibles, bien qu'ils soient d'accord sur les 22 premiers caractères. Ce simple fait signifie que les mots composés entièrement de caractères dans cet ensemble qui se chevauchent seront "corrects" dans l'un ou l'autre encodage. Lors de la vérification de l'encodage d'une liste de mots triée (comme dans un dictionnaire), il est sage de regarder au-delà des premiers mots pour ne pas tomber dans cette ambiguïté.

3.4.2. Unicode

Unicode est la norme de facto actuelle pour coder un grand nombre de langues et de scripts simultanément. Unicode a été conçu à l'origine pour utiliser deux octets d'informations (pour coder 65536 symboles uniques) et a été étendu depuis pour couvrir plus d'un million de symboles uniques. Pour l'arabe, Unicode prend en charge un jeu de caractères arabe étendu. Il donne également aux lettres arabes et aux ligatures des adresses uniques sous ce qu'elles appellent les diagrammes A et B de formulaires de présentation. L'inverse peut être avec perte. Bien qu'Unicode constitue une solution importante pour représenter l'ensemble de scripts en arabe étendu, il introduit de nouveaux défis. En particulier, il fournit différentes façons pour représenter le même symbole. Par exemple, les chiffres indo-arabes sont tous reproduits. De même, certaines lettres ont des formes qui ne peuvent pas être distinguées facilement. Par exemple, 8 (U + 0643) Arabe k et _ (U + 06A9) Persan k, qui ont initialement une forme similaire : \. est tapé sur un clavier Persan enfin. La présence de graphiques de formulaire de présentation permet un codage allographique incorrect qui peut ne pas être facilement détectable visuellement. Tous ces cas rendent difficile l'appariement de chaînes de texte identiques à l'écran, bien qu'ils soient encodés différemment.

3.5. Morphologie Arabe

La morphologie joue un rôle central dans le travail sur le TAL arabe en raison de ses interactions importantes avec l'orthographe et la syntaxe. La riche morphologie de l'arabe est peut-être la plus étudiée et la plus écrite sur l'aspect de l'arabe. En conséquence, il existe une profusion de terminologies, dont certaines incohérentes, qui peuvent intimider et embrouiller les nouveaux chercheurs.

La morphologie est l'étude de la structure interne des mots. Nous distinguons deux types d'approches de la morphologie : la morphologie basée sur la forme et la morphologie fonctionnelle. La morphologie basée sur la forme concerne la forme des unités composant un mot, leurs interactions les unes avec les autres et leur relation avec la forme générale du mot. En revanche, la morphologie fonctionnelle concerne la fonction des unités à l'intérieur du mot et comment elles affectent son comportement global syntaxiquement et sémantiquement.

Un graphe des différents termes morphologiques est présenté dans la Figure 12.

3.5.1. Morphologie à base de forme

Un concept central dans la morphologie basée sur la forme est le morphème, la plus petite unité significative dans une langue. Un trait distinctif de la morphologie arabe (en fait, sémitique) est la présence de morphèmes templatiques en plus des morphèmes concaténatifs. Les morphèmes concaténatifs participent à la formation du mot via un processus de concaténation séquentielle, tandis que les morphèmes templatiques sont entrelacés (interdigités, fusionnés).

3.5.2. Morphologie fonctionnelle

En morphologie fonctionnelle, nous étudions les mots en fonction de leur comportement morphosyntaxique et morphosémantique, par opposition à la forme des morphèmes à partir desquels ils sont construits. Nous distinguons trois opérations fonctionnelles : la dérivation, la flexion et la clitisation. La distinction entre ces trois opérations en arabe est similaire à celle des autres langues. Ce n'est pas surprenant puisque la morphologie fonctionnelle tend à être une manière plus indépendante du langage des mots.

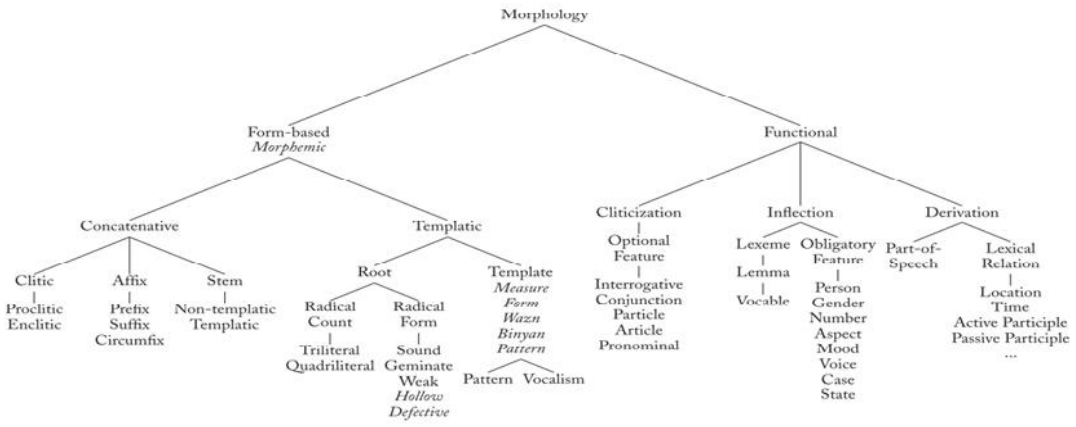

Figure 11: Graphe des relations morphologiques

3.6. Les tâches de la morphologie computationnelle

3.6.1. Tokenization

La tâche commune en TAL est que la segmentation des mots arabes par la déclinaison et la normalisation orthographique réductive est utile pour de nombreuses applications, telles que la modélisation du langage (LM), l'IR et le MT statistique (SMT). La titulation et la normalisation réduisent la rareté et les mots hors vocabulaire (OOV).

3.6.2. Etiquetage morphosyntaxique (POS TAGGING)

L'étiquetage morphosyntaxique (Part-Of-Speech (POS) tagging) consiste à attribuer une balise morphosyntaxique contextuellement appropriée à chaque mot d'une phrase. Les étiquettes sont sélectionnées à partir d'un jeu d'étiquettes qui, en principe, doit être bien défini et complet. En raison de sa morphologie riche, les jeux d'étiquettes arabes peuvent être très volumineux. De nombreux chercheurs travaillant en arabe préfèrent travailler avec des ensembles réduits plus petits. La taille et granularité d'un jeu d'étiquettes arabe peuvent varier énormément. D'un côté, la grammaire arabe traditionnelle est classée en : nom, verbe et particule. C'est une classification très grossière, qui n'est souvent pas utilisée de manière informatique. À l'autre extrémité, l'étiquette de Buckwalter à base de morphèmes arabes peut atteindre plus de 330000 étiquettes. La taille du jeu de balises interagit également avec le fait que le texte soit sous forme de Token ou non (et dans quel schéma de Tokenization). En principe, la balise POS d'un mot non reconnu est égale à la mise en chaîne des balises POS de ses Token.

3.6.3. Présentation des analyseurs morphologiques

Le domaine du TAL arabe a reçu beaucoup de contributions au cours des dernières décennies. De nombreux analyseurs traitent la morphologie-riche dans l'Arabe en MSA (Modern Texte Standard) (Atwell 2015). Dans le Tableau 2, on résume quelques analyseurs morphologiques de la langue arabe.

Analyseur morphologique	Description	Sortie
BAMA (Buckwalter 2004a).	Un analyseur morphologique arabe librement disponible, basé sur Perl et largement connu, par Tim Buckwalter.	Balise POS, mot voyelle et stem. Le jeu de balises de Buckwalter contient environ 70 sous-étiquettes de base, et peut être combiné pour former des balises plus complexes, telles que IV_PASS, qui signifie ″verbe passif imparfait″. Ces étiquettes incluent des caractéristiques de verbes comme la personne, la voix, l'humeur, l'aspect et le sujet, comme le sexe et le nombre. Il comprend également des fonctionnalités nominales telles que le sexe, le nombre, le cas et l'état. BAMA fournit une liste de différentes analyses

		sans aucune ambiguïté.
Mada (Pasha et al. 2014)	Une boîte à outils librement disponible qui numérise, balise-pos, balise et introduit une entrée en arabe brut. Cette boîte à outils, son successeur MADAMIRA, désambiguïsera les analyses en montrant la probabilité de chaque analyse.	Balise POS, mot voyelle, stem et lemme. Le jeu d'étiquettes de sortie peut être l'un des quatre jeux de POS TAG différents : ALMORGEANA, CATiB, POS: PENN, Penn ATB ou Buckwalter. Les caractéristiques des verbes comme la personne, la voix, l'humeur, l'aspect et son sujet, comme le genre et le nombre, sont explicitement fournies. Même chose pour les caractéristiques nominales, telles que le sexe, le nombre, la casse et l'état. MADA fournit une liste de différentes analyses avec une probabilité : le plus élevé est le plus probable.
MADAMIRA (Pasha et al. 2014)	est le successeur de MADA basé sur Java qui combine les outils MADA et AMIRA. Il ajoute quelques aspects de l'outil AMIRA.	En plus de la sortie de MADA, la base des morceaux de phrase et des entités nommées peut être fournie.
AlKhalil (Boudlal et al. 2010)	"un analyseur morphosyntaxique" de MSA combinant une approche basée sur des règles et une analyse de table.	AlKhalil est différent car toute sortie est une forme de tableau en arabe qui décrit l'analyse morphologique de chaque mot. Le tableau comporte des colonnes de POS-tags, de préfixes, de suffixes, de motifs, de racines, de mots et de voyelles. Les caractéristiques des verbes comme la voix, la transitivité et l'aspect sont extractibles. Cependant, l'humeur et la personne ne sont pas explicitement fournies, ni son sujet, s'il s'agit d'un suffixe. Les caractéristiques nominales sont également extractibles. En outre, AlKhalil fournit la nature du nom, de la racine du mot et de la forme du verbe.
ATKS (Microsoft	est un service Web de composants NLP	Comme les tags Buckwalter, ATKS fournit des tags complexes qui englobent

2018)	ciblant la langue arabe. Il comprend un analyseur morphologique à part entière (Sarf) et une partie du discours (POS) Tagger.	les fonctions nominales et verbales. Toutes les fonctionnalités sont extractibles des balises POS. Sarf fournit une liste de fonctionnalités, telles que stem, root, pattern et Token discrétisé.
Les outils TAL de Stanford (Toutanova et al. 2013)	un logiciel open-source en Java qui possède un segmenteur, un pos-tagger et un analyseur de texte arabe.	La sortie de l'analyseur et du postagger de Stanford est Bies tagset qui est utilisée pour l'arabe Penn Treebank. Cet ensemble de balises est linguistiquement grossiée (Habash 2010) et par conséquent, de nombreuses fonctionnalités sont manquantes. Les fonctionnalités extractibles sont l'aspect (sauf s'il est passif, car les verbes parfaits et imparfaits partagent la même étiquette), le numéro (singulier ou pluriel uniquement) et la voix.

Tableau 2: Exemple d'analyseurs morphologiques de la langue arabe

4. Le Web sémantique Arabe

Le Web sémantique basé sur la langue arabe souffre d'un manque de ressource en ce qui concerne la technologie, les outils et les applications. Il y a beaucoup de langues que la plupart des entreprises du monde entier ne prennent pas en compte, et l'arabe est l'une d'eux. En outre, il n'y a pas de société arabe qui se préoccupe de la fabrication de dispositifs qui traitent directement la langue arabe. De plus, il n'est pas inclus dans l'intérêt ou les objectifs des pays arabes de construire ou d'appuyer des centres de recherche pour aider la langue et la population arabes à suivre l'évolution de la technologie.

En outre, la majorité des outils et applications du Web sémantique ne supportent pas d'une façon complète la langue arabe. Cependant, il existe quelques articles publiés qui traitent ce problème, et la plupart de ces tentatives montrent seulement pourquoi la langue arabe n'est pas prise en charge par les outils Web sémantiques.

Comme le montre le tableau 6, les outils de gestion des ontologies pour la langue arabe présentent des faiblesses. Il est donc nécessaire de développer des outils de gestion du vocabulaire des ontologies arabes pour le développement du Web sémantique arabe (Beseiso et al. 2010).

Outil	Ontologie Arabe RDF	Ontologie Arabe OWL	Requête en langue arabe	Description
Protege	Compatible	Compatibilité limitée	Compatibilité limitée	Editeur d'Ontologie
Jena	Compatible	Compatible	Compatibilité limitée	raisonnement et processeur
Sesame	Compatibilité limitée	Compatibilité limitée	Non Compatible	Base de données RDF
KAON2	Non Compatible	Non Compatible	Non Compatible	Raisonneur et gestion d'ontologie

Tableau 3: Exemple des outils du Web sémantique et leurs compatibilités avec la langue arabe

Le tableau 6 montre à quel point certains outils Web sémantique prennent en charge la langue arabe. La première colonne du tableau 6 correspond au nom de l'outil. Les colonnes "ontologie arabe RDF" et "ontologie arabe OWL" indiquent si l'outil prend en charge ces deux langages (RDF et OWL) pour exprimer une ontologie arabe. La colonne "Requête en langue arabe" indique si l'outil permet d'interroger l'ontologie à l'aide d'un langage de requête en langue arabe. La colonne "Description" indique la fonction principale de l'outil

4.1. Recherches sur les ontologies et le Web sémantique arabe

Il existe plusieurs études menées par beaucoup de chercheurs qui associent les valeurs arabes et sémantiques. Un état de l'art sur de nombreuses tentatives de recherche ayant présenté des données sur la langue arabe et le Web sémantique pourrait aider davantage à mieux comprendre le problème.

Dans le domaine juridique, S. Zaidi et al. dans (Zaidi et al. 2005) a présenté un outil multilingue basé sur le Web pour la recherche d'information en arabe basée sur une ontologie de domaine juridique, dans le but d'améliorer à la fois le Rappel et la Précision de la recherche. L'outil tente de minimiser le niveau de bruit dans les résultats de recherche sur le domaine juridique à l'aide d'une extension de requête basée sur une ontologie. L'ontologie arabe juridique est construite manuellement par Protégé en utilisant une stratégie descendante dans laquelle les concepts hiérarchiques sont liés via une relation is-a. Pour la population ontologique, les articles arabes des Nations Unies sont utilisés, de même que certains journaux arabes.

Le domaine agricole en arabe est conceptualisé dans d'autres ontologies de domaine présentées à (Hazman et al. 2009). Cet article propose un système qui automatise le processus de construction d'une ontologie taxonomique à l'aide de documents Web spécifiques à un domaine semi-structuré. Le système construit une ontologie de domaine agricole en utilisant un ensemble de 180 documents d'extension en arabe avec 3817 titres HTML et 30 concepts de semences représentant les principaux concepts de la production agricole. L'ontologie est construite en utilisant deux approches complémentaires. La première consiste à utiliser la structure de phrases qui apparaît dans les en-têtes HTML des documents utilisés. La seconde consiste à utiliser la structure hiérarchique des titres HTML pour identifier de nouveaux concepts et leurs relations taxonomiques entre les concepts de base et entre eux même.

Belkedim et Meziane dans (Belkredim and Meziane 2008) ont abordé le développement de l'ontologie arabe en utilisant des verbes et des racines, car les verbes sont classés par leurs règles de dérivation en arabe à partir de racines et 85% des mots arabes proviennent de racines tri-littérales.

Outre que les auteurs n'ont pas présenté de modèle implémenté pour étayer leur hypothèse, l'utilisation des racines dérivées comme base pour construire une ontologie est imprécise puisque ces mots dérivés, bien qu'ils aient la même signification centrale, pourraient être classés dans différentes classes. Le mot 'علم', par exemple, n'appartient pas à la classe 'Human' qui encapsule le mot 'معلّم', bien qu'ils aient la même racine (Al-Zoghby et al. 2013).

Saleh et Al-Khalifa dans (Saleh and Al-Khalifa 2009) ont proposé un modèle supplémentaire pour construire et utiliser des ontologies arabes spécifiques à un domaine. Cette fois, c'était pour le "domaine de localisation". C'est une recherche qui présente un outil d'annotation sémantique en arabe appelé AraTation pour annoter sémantiquement le contenu d'Arabe News Documents Web.

Pour le domaine islamique, plusieurs recherches ont présenté une représentation ontologique du savoir islamique. Par exemple, la recherche (Al-Khalifa et al. 2009) représente le champ des «termes d'opposition» dans le Saint Coran, tandis que (Beseiso et al. 2011) représente le champ «noms temporels». La recherche (Sharaf et al. 2010), quant à elle, représente un graphe conceptuel comme une ontologie plus large pour le Saint Coran. Les auteurs de (Salim et al. 2010) considèrent l'ontologie multilingue pour le portail islamique.

D'autre part, (Moawad et al. 2010) est un autre travail qui a développé une ontologie arabe dans le domaine de la technologie informatique. Cependant, il traite de la langue traditionnelle ou classique plutôt que de la langue moderne. En outre, l'ontologie présentée est beaucoup plus simple, et ne contient pas ce nombre de classes.

4.2. Le projet Dbpedia

Wikipedia est devenue l'une des sources de connaissances centrales de l'humanité et est maintenue par des milliers de contributeurs. Les articles de Wikipedia sont principalement composés de texte en langage naturel, mais contiennent également différents types d'informations structurées, telles que les modèles infobox, les informations de catégorisation, les images, les coordonnées géographiques et les liens vers des pages Web externes.

Le projet DBpedia (Bizer et al. 2009) extrait divers types d'informations structurées issues des éditions Wikipedia dans plusieurs langues via un framework d'extraction open source. Il combine toutes ces informations dans une base de connaissances multilingue multi-domaine. Pour chaque page de Wikipedia, un identificateur de ressource uniforme (URI) est créé dans DBpedia pour identifier une entité ou un concept décrit par la page Wikipedia correspondante. Au cours du processus d'extraction, les informations structurées provenant du wiki telles que les champs infobox, les catégories et les liens de page sont extraites en tant que triplets RDF et ajoutées à la base de connaissances en tant que propriétés de l'URI correspondant.

4.2.1. L’ontologie Dbpedia

L'ontologie DBpedia organise les connaissances sur Wikipedia en 320 classes qui forment une hiérarchie de subsomption et sont décrites par 1650 propriétés différentes. DBpedia gagne sa publicité grâce à un grand nombre de ressources et de jeux de données qui lui sont associés, et qui forment les données ouvertes liées (LOD). La version de DBpedia disponible au moment de la rédaction de cette thèse est la version de DBpedia d’octobre 2016. La version anglaise de la base de connaissances DBpedia décrit actuellement 6,6 millions d’entités, dont 4,9 millions de résumés, 1,9 millions de coordonnées géographiques et 1,7 millions de représentations. Au total, 5,5 millions de ressources sont classées dans une ontologie cohérente, soit 1,5 millions de personnes, 840000 places, 496000 œuvres (dont 139000 albums de musique, 111000 films et 21000 jeux vidéo), 286000 entreprises et 55000 établissements d'enseignement, 306 milles espèces, 58 milles plantes et 6000 maladies. Le nombre total de ressources en anglais DBpedia est de 18 millions. Outre les ressources est de 6,6 millions, il comprend 1,7 millions de concepts (catégories), 7,7 millions de pages de redirection, 269 pages de désambiguïsation et 1,7 million de nœuds intermédiaires (DBpedia 2018).

L'alignement entre les infoboxes de Wikipedia et l'ontologie se fait via des mappages fournis par la communauté. Ces mappages permettent de normaliser la variation de nom dans les propriétés et les classes. Les hétérogénéités du système d'infoboxe de Wikipedia, comme l'utilisation de différentes infoboxes pour le même type d'entité (classe) ou l'utilisation de noms de propriété différents pour la même propriété, peuvent être obtenues de cette manière. Par exemple, «date of birth» et «birth date» sont tous deux mappés sur la même propriété birthDate et les infoboxes «Infobox Person» et «Infobox FoundingPerson» ont été mappées par la communauté DBpedia vers la classe Person. Les mappages DBpedia existent actuellement pour 23 langues. Ce qui signifie que d’autres propriétés infobox, telles que «data de nascimento», «Geburtstag», « تاريخ الميلاد» - date de naissance dans le portugais, l'allemand et l’arabe, respectivement, sont également associées à la date de naissance de l'identifiant global. Cela signifie que les informations de toutes les versions linguistiques de DBpedia peuvent être fusionnées. Les bases de connaissances pour les langues les plus petites peuvent donc être enrichies de connaissances provenant de sources plus importantes, telles que l'édition anglaise. À l'inverse, les éditions DBpedia les plus importantes peuvent bénéficier de connaissances plus spécialisées issues des éditions localisées (Tacchini et al. 2009).

4.2.2. Le chapitre Arabe de Dbpedia

Wikipedia est considéré comme un environnement de collaboration, de création, d'édition et de publication de contenus provenant de divers contributeurs dans le monde entier (Auer and Lehmann 2007). Wikipedia représente ses données dans un format simple appelé texte wiki, simple à modifier et à éditer. En plus de ça, il peut décrire à la fois le contenu structuré et non structuré. Le modèle est considéré comme un moyen de représentation des articles dans Wikipedia. L'article Wikipedia commence généralement par un court paragraphe décrivant le sujet. Un exemple célèbre de gabarit dans Wikipedia est une infobox qui est une boîte formatée recueillant les principaux points dans l'article de Wikipedia et qui est placée derrière le bref résumé de la page.

```
صندوق معلومات شخص}}
-الاسم                |
=الاسم الأصلي         |
=الصورة               |
- تعليق الصوره|
<-- {{تاريخ الميلاد والعمر|السنة|الشهر|اليوم}} --!> = تاريخ الولادة|
=مكان الولادة         |
<-- {{mf-yes|تاريخ الوفاه والعمر|الوفاه|1|1|الميلاد|1|1}} --!> - تاريخ الوفاه|
=مكان الوفاة          |
=سبب الوفاة           |
=مكان الدفن           |
=الإقامة              |
<-- علم ديكو|اسم البلد الجنسية}} --!> =الجنسية      |
=المواطنة             |
=تعليم                |
=المدرسة الأم         |
=المهنة               |
= سنوات النشاط|
=سبب الشهرة           |
=الديانة              |
=الزوج                |
=الجوائز              |
-التوقيع              |
=الموقع الرسمي|
                      }}
```

Listing 6. Template de Mappage pour infobox pays « معلومات بلد »

Il est considéré comme l'élément d'information le plus important de Wikipédia extrait de DBpedia. Au-delà de l'infobox produit, il existe une syntaxe décrivant les fonctionnalités apparaissant à l'utilisateur dans l'infobox. Infobox se compose de paires qui représentent le nom de l'attribut et sa valeur. Tous les attributs n'ont pas de valeurs, et cela peut revenir à plusieurs raisons, dont l'une est liée à l'indisponibilité de la valeur de l'attribut pour l'auteur de la base de données Wikipedia qui peut ne pas être le même auteur de l'article.

Dans l'édition arabe de Wikipedia, l'infobox est placée en haut à gauche de la page Wikipedia, à l'opposé de l'édition anglaise. Les attributs peuvent être écrits en arabe ou en anglais, mais il est préférable qu'ils soient écrits en arabe, leurs valeurs d'attribut doivent être écrites en arabe. DBpedia utilise des noms d'article pour créer des identificateurs. Ces noms sont extraits de l'URL de l'article Wikipedia via http://ar.wikipedia.org/wiki/Name et l'URI de cette ressource dans DBpedia est devenu http://ar.dbpedia.org/resource/Name.

Il existe deux méthodes d'extraction, généralement utilisées dans DBpedia, décrites en détail dans (Bizer et al. 2009). Elles sont l'extraction d'Infobox générique et l'extraction d'Infobox basée sur le mappage. Dans l'extraction d'Infobox générique, tous les attributs et leurs valeurs dans les infoboxes de Wikipedia sont extraits dans l'article suivant : l'article sera le sujet, l'attribut infobox concaténé avec l'espace de noms de http://dbpedia.org/propenty/ sera un prédicat, tandis que la valeur de l'attribut infobox dans Wikipedia est considérée comme un objet.

D'autre part, l'extraction d'infobox basée sur le mappage est utilisée pour couvrir les lacunes de la première méthode et elle dépend de la mise en correspondance des modèles Wikipedia avec une ontologie créée pour collecter les propriétés les plus courantes utilisées dans Wikipedia.

5. Conclusion

Dans ce chapitre, nous avons discuté le Web sémantique et la langue arabe. L'évolution du Web depuis sa création a connu un succès sans précédent. Il est devenu la technologie informatique la plus utilisée au quotidien. La prochaine génération sera le Web sémantique qui est actuellement en plain évolution, la langue arabe doit intégrer cette technologie pour garder son existence dans le monde numérique. Le chapitre DBpedia est un projet prometteur dans lequel la plus grande encyclopédie numérique est liée à une ontologie universelle. Le chapitre arabe est le chapitre le plus récent ajouté à DBpedia ; notre langue arabe a besoin d'un effort énorme pour être à jours avec les technologies Web et informatique, actuelles et futures.

Chapitre 4 : Etat de l'art sur les systèmes question-réponse

1. Introduction

Les systèmes question-réponse permettent à l'utilisateur de poser une question en langage naturel (NL) et de retourner la bonne réponse à sa question au lieu d'un ensemble de documents jugés pertinents, comme c'est le cas pour les moteurs de recherche. Cependant, pour les systèmes question-réponse visant les textes et les documents Web, la structure des informations requises affecte la précision de ces systèmes. Les SQR sont plus efficaces pour interagir avec des données structurées.

Dans ce chapitre on présente un état de l'art sur les SQR en détaillant les trois classes des systèmes, ainsi que les travaux relatifs en différentes architectures. La section suivante introduit les interfaces de langage naturel pour les bases de données. La section 3 présente les systèmes question-réponse textuels. La section 4 présente les systèmes question-réponse pour les données liées. Dans la section 5, nous présentons les performances des deux types de SQR: SQR basé sur l'ontologie et SQR basé sur le texte. On termine dans la section 6 par une conclusion.

2. Interface de langage naturel pour les bases de données ILNBD

ILNBD (*Natural language Interface for Database*) est un système qui permet à un utilisateur simple d'accéder aux informations stockées dans des bases de données en formulant les requêtes exprimées en langage naturel, par exemple : en anglais, en français, en arabe, etc.

L'exemple suivant est un dialogue entre un utilisateur et LOQUI qui un ILNBD commerciale. Les réponses du système sont légèrement simplifiées (Binot et al. 1991).

```
> Who works on 3 projects?
B. Vandecapelle, C. Willems, D. Sedlock, J.L. Binot, L.
Debille, ...
> Which of them are project leaders?
D. Sedlock, J.L. Binot
> Documents describing their projects?
Bim Loqui: "The Loqui Nlidb", "Bim Loqui"
Mmi2: "Technical Annex"
> How many of these projects do not finish before 1994?
2
Bim Loqui, Mmi2
> Are they led by JLB or DS?
The former.
```

Listing 7. Exemple de dialogue entre un utilisateur et LOQUI

Les premiers ILNBD remonte à la fin des années 60 et début des années 70. Le ILNBD le plus connu dans cette période est LUNAR (Woods et al. 1972), qui est une interface de langage naturel pour les bases de données contenant des analyses chimiques des roches lunaires. LUNAR et les premiers ILNBD ont été construits en tenant compte d'une base de données particulière, et ne pourraient donc pas être facilement modifiés pour être utilisés avec différentes bases de données. Bien que la méthode de représentation interne utilisée dans

LUNAR a été étendue pour faciliter l'indépendance entre la base de données et d'autres modules (Woods 1968), la manière dont LUNAR a été utilisé était quelque peu spécifique aux besoins de ce projet.

À la fin des années 70, plusieurs autres ILNBD sont apparus, par exemple RENDEZVOUS (Codd 1974), LADDER (Hendrix et al. 1978), PHILIQA (Scha 1977), CHAT80 (Warren 1982). Bien que les grammaires sémantiques aient aidé à mettre en œuvre des systèmes aux caractéristiques impressionnantes, les systèmes résultants se sont révélés difficiles à porter sur des domaines d'application différents. Comme les chercheurs ont commencé à se concentrer sur les ILNBD portables, les grammaires sémantiques ont été progressivement abandonnées.

Au milieu des années 80, le domaine des ILNBD était un domaine de recherche très populaire, et de nombreux prototypes de systèmes étaient en cours d'implémentation. Une grande partie des recherches de cette époque a été consacrée aux problèmes de portabilité. Les ILNBD sont conçus pour être facilement configurable par les administrateurs de la base de données.

Bien que certains ILNBD développés au milieu des années 80 aient démontré des caractéristiques impressionnantes dans certains domaines d'application, les ILNBD n'ont pas réussi à être commercialisés. Au cours des dernières années, le nombre d'articles publiés par an sur NlLIDB a considérablement diminué. Pourtant, les ILNBD continuent d'évoluer, en adoptant des progrès dans le domaine du traitement automatique du langage naturel.

2.1. L'architecture d'un ILNBD

Les chercheures informatiques ont divisé le problème de ILNBD en deux sous-composants :

1. composant linguistique,
2. et composant base de données.

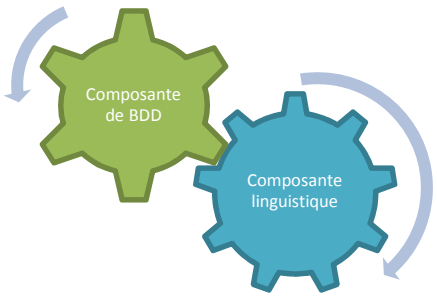

Figure 12: Composants des ILNBD

a. Composant linguistique

Ce composant est responsable de la traduction de la question exprimée en langage naturel vers une requête formelle, et de la génération d'une réponse en langage naturel basée sur les résultats d'exécution de la requête dans la base de données. Le processus de conversion est basé sur les techniques de traitement automatique du langage naturel. Les techniques utilisées reposent sur l'analyse lexicale, l'analyse syntaxique et la reconnaissance des entités nommées.

b. Composant de base de données

Le composant de BDD effectue des fonctions traditionnelles de gestion de base de données. Un lexique est une table utilisée pour mapper les mots générés par le composant linguistique de l'entrée naturelle sur les objets formels (noms de relations, noms d'attributs, etc.) de la base de données.

Un générateur de langage naturel prend la réponse formelle comme entrée, et inspecte l'arbre d'analyse syntaxique afin de générer une réponse adéquate en langage naturel. (Neelu Nihalani 2011).

2.2. **Avantages et inconvénients des ILNBD**

Les ILNBD, représentent une alternative efficace contre les solutions traditionnelles pour la recherche d'information dans une base de données, comme les langages de requêtes formelles, les interfaces des formulaires et les interfaces graphiques.

L'accès à l'information stockée dans une base de données utilise traditionnellement un langage de requêtes tel que SQL. L'exemple montré dans la Figure 15 montre une base de données relationnelle :

Table-etudiants

Etudiant	*Departement*	*e-mail*
Mohamed	*mathématique*	*mohamed@gmail.com*
Ali	*informatique*	*ali@gmail.com*

Table-departements

Département	*Chef-depatement*	*e-mail*
mathématique	*djelloul*	*mohamed@gmail.com*
informatique	*benamar*	*ali@gmail.com*

Figure 13: Tables de BDD relationnelle

Dans la table des étudiants, on trouve le nom de l'étudiant, son département et son e-mail. Dans la table des départements, on trouve la désignation du département, le chef de département et l'université. Si on veut afficher pour chaque chef de département, la liste des étudiants correspondants, on écrit la requête SQL suivante :

```
SELECT table-etudiants.etudiant, table-departement.chef-departement
FROM table-etudiants, table-departement
WHERE table-etudiants.department = table-departement.department
```

Listing 8. Exemple de requête SQL

Dans une interface de formulaire, des formes prédéfinies constituées de champs sont utilisées. L'utilisateur remplit les champs du formulaire, et le système produit une réponse par l'interrogation de la BDD avec une requête.

Dans une interface graphique, l'utilisateur sélectionne premièrement les tables qu'il veut utiliser. Ils sont affichés avec leurs attributs sur l'écran. L'utilisateur peut alors remplir les emplacements d'attributs. En tapant sur le clavier, il peut imposer des restrictions sur les attributs avec l'utilisation de la souris ou le menu disponible.

Les avantages et les inconvénients suivants sont souvent mentionnés dans la littérature :

2.2.1. Avantages des ILNBD

a. Pas de langage artificiel

L'utilisateur ordinaire n'est pas obligé d'apprendre un langage de communication artificiel. Les langages de requête formels comme SQL sont difficiles à apprendre et à maîtriser, du moins par des non-spécialistes en informatique.

b. Simple et facile à utiliser

Les applications traditionnelles de base de données utilisent un langage de requête ou un certain formulaire conçu pour afficher la requête. Un formulaire peut contenir plusieurs entrées (champs, boîtes de défilement, listes déroulantes, boutons radio, etc.) en fonction de la capacité du formulaire. Dans le cas d'un langage de requête, il peut être nécessaire d'exprimer une question à l'aide de plusieurs instructions qui contiennent une ou plusieurs sous-requêtes avec des opérations conjointes en tant que connecteur. Cependant, un système ILNBD ne nécessite qu'une seule entrée (une question en langage naturel).

c. Efficace pour certaines questions

Il existe certains types de questions (par exemple, les questions de négation, ou de quantification) qui peuvent être facilement exprimées en langage naturel, mais qui semblent difficiles (ou du moins fastidieuses) à exprimer à l'aide d'interfaces graphiques ou requêtes formelles.

d. Tolérance à l'erreur

La plupart des systèmes ILNBD offrent des tolérances aux erreurs grammaticales mineures, tandis que dans une application de base de données, le lexique des mots et la syntaxe des requêtes formelles devraient être corrects, et toute erreur entraînerait automatiquement le rejet de la requête par le SGBD.

2.2.2. Inconvénients des ILNBD

a. La couverture linguistique

Actuellement, tous les systèmes ILNBD 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 ILNBD ne peuvent pas répondre à certaines questions appartenant à leurs propres sous-ensembles. Ce n'est pas le cas dans un langage formel.

b. Échecs linguistiques ou conceptuels

Dans le cas de défaillance du système ILNBD, il arrive souvent que le système ne fournisse aucune explication sur les raisons de l'échec. Certains utilisateurs peuvent essayer de reformuler la question ou simplement laisser la question sans réponse. La plupart du temps, il appartient aux utilisateurs de déterminer les causes de l'erreur.

c. Surestimation

Les gens peuvent surestimer un système ILNBD à traiter un langage naturel. Ils peuvent supposer que le système est intelligent. Par conséquent, plutôt que de poser des questions précises sur une base de données, ils peuvent poser des questions complexes.

2.3. Travaux sur les ILNBD

LUNAR, introduit en 1971, est un ILNBD à domaine spécifique qui répond à des questions sur des analyses chimiques des roches lunaires (Woods et al. 1972).

RENDEZVOUS est une interface de langage naturel où la requête de l'utilisateur est donnée sous une boite de dialogue et une terminologie spéciale pour clarifier la question au système (Codd 1974).

PHILIQA permet une compréhension sémantique de la question de l'utilisateur en trois étapes : "langage formel anglais", "model de langage standard" et "langage de BDD". Il est aussi connu sous le nom de Philips Question Answering System. (Scha 1977)

CHAT-80 est l'un des systèmes les plus référenciés dans les années 80. CHAR-80 est implémenté en PROLOG. La base de données de CHAT-80 est composée de faits (océans, mers majeurs, rivières majeurs, villes majeurs) sur 150 pays du monde et un petit ensemble du vocabulaire anglais. (Warren 1982)

LADDER utilise une base de données distribuée. C'est une interface de langage naturel dans laquelle la question de l'utilisateur est parsée en utilisant une grammaire syntaxique (Hendrix et al. 1978).

JANUS est une interface pour de multiples sources de données (BDD, système expert, dispositif graphique). L'hétérogénéité des sources de données est cachée. L'utilisateur ignore la source de données. (Resnik 1989)

MASQUE/SQL est une interface de langage naturel portable et semi-configurable pour les bases de données SQL (Androutsopoulos et al. 1993).

2.3.1. Les ILNBD les plus récents

NALIX une interface de langage naturel pour les bases de données XML. L'idée principale de ce système consiste à utiliser MQF (Meaningful Query Focus) pour trouver les correspondances entre les mots clés de la requête et les éléments XML. Il n'est pas nécessaire de faire un mappage (Li et al. 2006). Une brève description du système NLWIDB est donnée par les auteurs à travers l'exemple suivant : considérant une base de données appelée UNIVERSITY qui a été créée en utilisant MySQL pour une université. Au sein de la base de données UNIVERSITY, NALIX crées plusieurs tables qui sont correctement normalisées.

Maintenant, si l'utilisateur souhaite accéder aux données de la table dans la base de données, il doit être techniquement compétent dans le langage SQL pour formuler une requête pour la base de données UNIVERSITY. NALIX facilite les choses, et permet à l'utilisateur final d'accéder aux tables avec un langage naturel. Prenons un exemple : supposons que nous voulons voir des informations, telles que l'année d'ouverture du département, et le code du département dont le nom du département est égal à "Département mathématiques et informatique" de la table ″Département″ de la base de données UNIVERSITY. Nous utilisons donc la requête SQL suivante :

SELECT year-of-opening-of-department, code-of-department

FROM Department

WHERE department-name = 'Département mathématiques et informatique'

Mais une personne, qui ne sait pas la syntaxe de la base de données MySQL, ne sera pas en mesure d'accéder à la base de données UNIVERSITY à moins qu'il connait bien la syntaxe SQL et lance une requête sur la base de données. NALIX permet l'interroger la base de données en utilisant le langage naturel. Cet accès à la base de données sera beaucoup plus simple. Donc, la requête SQL sera écrite sur une interface Web en langage naturel de la façon suivante : Quelle est l'année d'ouverture du département et le code du département dont le nom du département est égal à "Département mathématiques et informatique" ?

PRECISE est un système développé à l'université Washington par Ana-Maria Popescu, Alex Armanasu, Oren Etzioni, David Ko, and Alexander Yates en 2004. La source de données est une base de données relationnelle qui utilise SQL comme langage de requête. Les auteurs ont introduit la notion du traitement sémantique des phrases qui sont des phrases translatables à une unique interprétation sémantique par l'analyse morphologique et la contrainte sémantique. PRECISE a été évalué sur deux domaines de bases de données. Le premier est ATIS domaine, qui consiste à poser des questions sur le voyage aérien avec une précision de 95%. Le deuxième domaine est GEOQUERY. Ce domaine contient des informations sur la géographie des USA. PRICES enregistre une précision de 100%. (Popescu et al. 2004)

WASP (Word Alignment-based Semantic Parsing) est un système développé à l'université de Texas, Austin par Yuk Wah Wong. Le système est conçu pour répondre à un objectif plus large de construire une représentation complète, formelle, symbolique et significative d'une phrase en langage naturel. Une logique des prédicats en PROLOG est utilisée comme un langage de requête. WASP utilise l'apprentissage pour construire un parseur sémantique, il utilise un corpus des questions en langage naturel annoté avec le langage de requêtes formelles. Il ne demande pas des connaissances préalables sur la syntaxe parce que l'apprentissage complet se fait avec une machine de translation statistique. WASP a été évalué en utilisant le domaine GEOQUERY, le même que PRECISE. Le corpus GEOQUERY est constitué de 880 questions pour l'apprentissage et 250 questions pour le test, regroupées sur une même base de test. Chaque base de test est divisée en 10 sous-ensembles équitables.

WASP a obtenu 86,14% de précision et 75% de rappel. Le système est évalué sur plusieurs langues : Anglais, Espagnol, Japonais et Türk. (Wong 2005)

3. 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 en exploitant les données liées.

Depuis la croissance constante du WS et l'émergence de la sémantique à grande échelle, la nécessité de créer des systèmes question-réponse sur les données liées QALD basés sur les ontologies est devenue très importante. Cette tendance a également été soutenue par des études d'utilisabilité (Kaufmann and Bernstein 2007), qui montrent que les utilisateurs occasionnels, généralement dépassés par la logique formelle du WS, préfèrent utiliser QALD pour interroger le WS. Les QALD combinent plusieurs sources de données structurées pour produire une réponse exacte à une question posée en langage naturel (Figure 16).

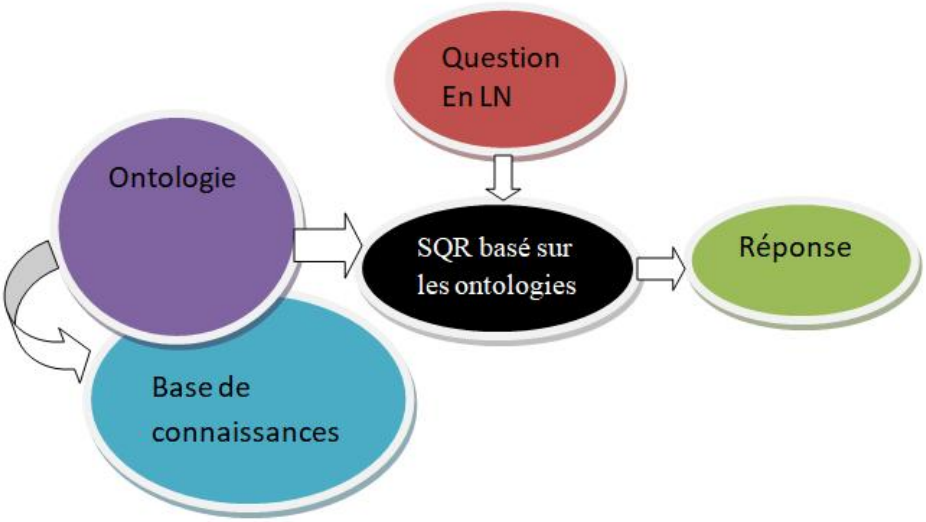

Figure 14: SQR basé sur les ontologies (entrée/sortie)

Ces dernières années, les QALD ont attiré beaucoup l'attention des chercheurs, où la puissance des données liées en tant que modèle de connaissances 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 couteuses du TAL, utilisées pour des domaines spécifiques, par des techniques peu profondes mais efficaces pour un 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 contrôlé ou guidé, langage naturel basé sur des patterns).

permettant 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).

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 connaissances. 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 d'entre eux respecte l'architecture globale basée sur trois modules : l'analyseur de question, formulation la requête SPARQL et le générateur de réponse (Figure 17).

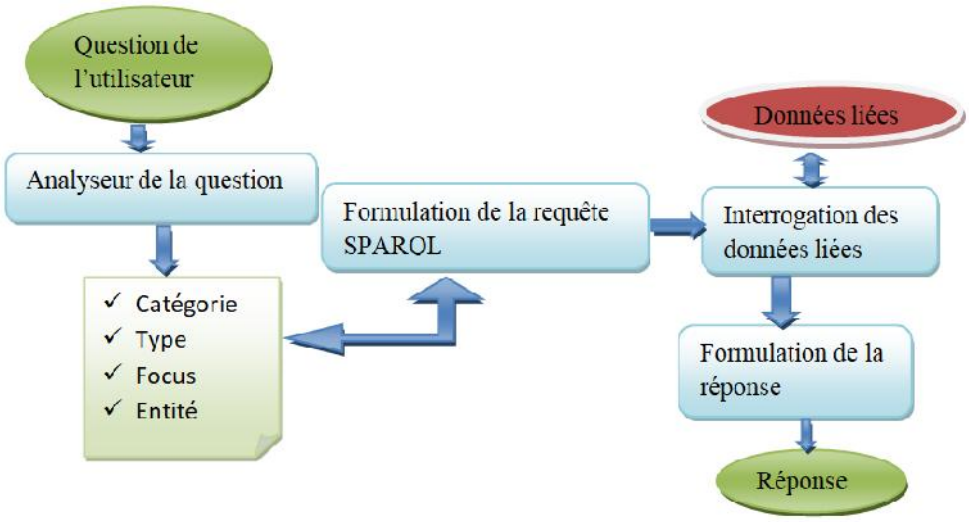

Figure 15: Architecture globale des SQRDL

Pour un 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 et les entités nommées. L'étape suivante consiste à utiliser les informations fournies 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, et le résultat est traité pour générer la réponse exacte à la question de l'utilisateur.

3.2. Questions en langage naturel et les données liées

La principale tâche des SQR est d'interpréter les informations de l'utilisateur exprimées 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 (b).

(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 de la république tchèque correspond à la ressource res:Czech_Republic, et 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 le 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, qui est relativement simple dans cet exemple particulier, très souvent, le processus est beaucoup plus compliqué. Dans la plupart des cas, cela implique deux défis: premièrement la correspondance des expressions du langage naturel aux éléments de vocabulaire utilisés par les données, en tenant compte des discordances lexicales et structurelles. Deuxièmement, en manipulant les variations de sens introduites par des expressions ambiguës et vagues, des expressions anaphoriques, etc. Regardons les deux défis suivants :

3.2.1. 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 le langage naturel 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 3 présente les différences lexicale et les différences structurelles entre les questions est les requêtes.

3.2.2. 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.

Difficultés	**Exemple**	**Description**
Différences de vocabulaire : la différence entre le vocabulaire du langage naturel et le vocabulaire utilisé par les données implique que les expressions utilisées par un utilisateur diffèrent sont souvent des étiquettes attachées aux données. Combler le fossé lexical qui en résulte est donc	Which Greek cities have more than 1 million inhabitants Select DISTANCT ? uri WHERE { ?uri rdf type dbo :city. ?uri dbo :contry res :Greece. ?uri dbo : populationTotal ?p. FILTER (?p>1000000)}	Il est plus ou moins simple de faire correspondre l'expression ″les villes″ à la classe ″dbo:Ville″ ayant la ville comme étiquette, que la correspondance des habitants à la propriété populationTotal. Ici, la similitude entre les deux existe seulement au niveau sémantique, mais pas au niveau des mots. De plus, la Grèce doit être appariée avec la propriété dbo:pays avec un

l'un des défis auxquels le SQR doit répondre.		objet fixe, à savoir Grèce.
Différences structurelles : elles sont dues au fait que la granularité conceptuelle du langage ne coïncide souvent pas avec celle du schéma qui repose sur un jeu de données. Soit le langage naturel est plus granulaire que les données (exemple 3) ou les données sont plus granulaires que le langage naturel (Exemple 4). En plus de mapper les expressions du langage naturel aux éléments de vocabulaire sous-jacents à un ensemble de données particulier, il existe des expressions qui ne correspondent à aucun élément de vocabulaire, mais qui ont plutôt une signification, indépendante de l'ensemble de données comme les quantificateurs (Exemple 5), les expressions de comparaison (Exemple 6).	When did Germany join the EU ? *SELECT DISTANCT ?date* *WHERE { res : Germany* *dbp :accessionedate ?date.* *}*	la structure des questions en LN suggère une relation qui relie deux entités, l'Allemagne et l'UE, tandis que la propriété requise est dbp:accessioneudate, reliant un pays à la date à laquelle il a rejoint l'UE.
	Who are the great-grandchild of Bruce lee ? *SELECT DISTINCT ?uri* *WHERE { res :Bruce-lee dbo: child ?c1.* *?C1 dbo :child ?c2.* *C2 dbo :child ?uri.* *}*	Il y a une expression en langage naturel arrière-petits-enfants qui correspond à une chaîne de propriétés constituée de trois fois la propriété dbo:enfant.
	Who produced the most films ? SELECT Distinct ?uri WHERE { ?x rdf :type dbo :film. ?x dbo :producer ?uri. } ORDER BY DSC(COUNT (?x)) LIMIT 1	La transformation des quantificateurs dans le langage naturel vers SPARQL.
	Which cities have more than three universities ? SELECT DISTINCT ?uri WHERE { ?x rdf :type dbo :University. ?x dbo :city ?uri. } HAVING (COUNT (?x) >3)	Les expressions de comparaissant sont souvent représentées par des formes de requêtes spécifiques.
	What is the second highest mountain on Earth ? SELECT DISTINCT ?uri WHERE { ?uri rdf :type dbo :Mountain. ?uri dbo :elevation ?x.} ORDER BY DESC (?x) OFFSET 1 LIMIT 1	Les cardinaux et superlatifs sont d'autres exemples où la structure de requête doit utiliser des agrégations pour exprimer la question de l'utilisateur.

Tableau 4: Difficultés de mappage des questions vers les requêtes

3.3. Travaux sur les SQRDL

Récemment, avec l'émergence du Web sémantique, beaucoup de travaux on vu le jour et la recherche dans ce domaine est prometteuse. Dans le Tableau 4, on présente des travaux sur les SQRDL.

AquaLog (Lopez et al. 2007) : Permet à l'utilisateur de choisir une ontologie, puis de poser des questions en langage naturel sur l'univers de discours couvert par l'ontologie.

PowerAqua (Lopez et al. 2012) : Un SQR qui interroge de multiples ressources dans le Web sémantique.

QACID (Ferrández et al. 2009) : Ce système s'appuie sur une ontologie, une collection de questions utilisateur et un mécanisme qui associe de nouvelles requêtes à un cluster de requêtes existant.

ORAKEL (Cimiano et al. 2007) : Un SQR qui transforme les questions factuelles en langage SPARQL ou logique. Cette transformation est évaluée par rapport à l'ontologie cible.

e-Librarian (Linckels and Meinel 2005) : Ce système analyse le sens de la question pour rechercher des ressources multimédias dans la base de connaissances.

GINSENG (Bernstein et al. 2005) : Ce système offre un vocabulaire contrôlé pour l'utilisateur via un vocabulaire fixe et des structures de phrases prédéfinies via des options de menu.

PANTO (Wang et al. 2007) : C'est une interface de langage naturel pour l'ontologie qui prend une question en entrée et exécute une requête SPARQL correspondante sur un modèle d'ontologie donnée.

QuestIO (Tablan et al. 2008) : Les questions sont traduites en requêtes formelles mais le système dépend de l'utilisation de nomenclatures initialisées pour l'ontologie de domaine.

FREyA (Damljanovic et al. 2011) : Fournit des améliorations concernant une compréhension plus profonde du sens sémantique d'une question.

QAKIS (Cabrio et al. 2012) : Ce système utilise des techniques du TAL pour l'appariement de fragments et de motifs textuels collectés automatiquement à partir de Wikipédia.

SPARQL2NL (Ngonga Ngomo et al. 2013) : Ce système fonctionne du côté de la conversion d'une requête SPARQL en langage naturel.

SWIP (Pradel et al. 2014) : Le traitement de la question est basé sur l'utilisation de la requête pivot, et la formalisation de cette requête pivot.

Pythia (Unger and Cimiano 2011) : Ce système utilise des ressources d'ontologie pour traduire la question vers le langage de requête.

(Al-Khalifa et al. 2009) : Le système proposé étend la requête de l'utilisateur dans les variantes de requête détendue à l'aide de la fonction de mappage de structures linguistiques vers des structures sémantiques compatibles avec l'ontologie.

SQUALL (Ferré 2013) : Ce système utilise un langage contrôlé pour la translation de la question en requête SPARQL.

LODQA (Ryu et al. 2014) : La question de l'utilisateur est transformée en une requête Template qui sera par la suite transformée automatiquement en requête SPARQL.

DeepQA IBM Watson's system (Fan et al. 2012): Ce système utilise des sources de données structurées et non structurées pour extraire et évaluer les réponses.

Xser (Xu et al. 2014) : Le système fonctionne en deux étapes : (1) en utilisant un analyseur sémantique pour l'analyse linguistique afin de détecter les structures de prédicat, (2) la requête est instanciée par rapport à la structure de la base de connaissances.

gAnswer (Huang and Zou 2013) : Le processus est conduit par un graphe, et fonctionne en deux étapes : (1) une analyse syntaxique de la question entraîne une structure sémantique de la question, (2) le graphique résultant est apparié à des triplets RDF. Une désambiguïsation est nécessaire pour faire correspondre les sous-graphes.

CASIA (He et al. 2014) : Un algorithme de réseau Markov Logic est utilisé pour entrainer un modèle commun, détecter des phrases et mapper des éléments sémantiques. Pour ces expressions, les éléments sémantiques sont regroupés dans un graphique.

Intui3 (Dima 2014) : Les techniques du TAL sont utilisées : la question est analysée syntaxiquement, fragmentée, et les entités nommées sont identifiées. Chaque fragment reçoit ensuite une ou plusieurs interprétations en fonction de son type et ses informations sémantiques et syntaxiques supplémentaires disponibles. En utilisant une combinaison de règles attachées pour chaque type d'interprétation du fragment, l'interprétation de la question est mappée vers une requête SPARQL correspondante.

ISOFT (Park et al. 2015) : Ce système transforme des questions en langage naturel en requêtes SPARQL. Il utilise une approche basée sur les modèles. Une analyse linguistique de la question, des modèles de requête sont déterminés, une recherche sur les concepts appropriés dans la base de connaissances est exécutée, en se basant sur la similarité lexicale et l'analyse sémantique.

Metafrastes (Embregts et al. 2014) : Le système récupère des informations à partir des bases de connaissances du Web sémantique, en utilisant des techniques du TAL, pour traduire la question en requête SPARQL.

4. Les systèmes question-réponse textuels

Les moteurs de recherche SE (search engine), par exemple google, yahou et altavista, sont devenus une habitude quotidienne pour les internautes. Ces SE rend le contenu textuel du Web exploitable par les utilisateurs qui introduisent un ensemble de mots clés. Il y a des situations où l'utilisateur a besoin des informations spécifiques, par exemple, la période de déroulement de l'histoire de ROMEO et JOULIET.

Une recherche avec les mots clés ROMEO et JUILIET peut nous rendre un document contenant la date de déroulement de l'histoire. Mais ça sera plus intéressent si on peut poser la question : dans quelle période l'histoire de ROMEO et JULIET s'est déroulée ? Et avoir la réponse : 13ième siècle.

Les systèmes question-réponse textuels utilisent un ensemble de techniques pour l'extraction d'une phrase ou un fragment de texte à partir d'un document ou d'une page Web, pour répondre exactement à une question. Les questions-réponses textuelles sur un domaine ouvert supposent que les questions sont naturelles et sans restriction par rapport au domaine.

Les techniques actuellement utilisées dans les systèmes question-réponse sont divisées en deux types (Harabagiu et al. 2003) : (1) les techniques de la recherche d'information qui localisent les réponses dans une large collection de documents ; et (2) les techniques de la compréhension automatique qui répond à un ensemble de questions sur un document donné. Ces deux techniques sont différentes, mais leurs combinaisons est désirable pour plus de performance dans les SQR.

Les SQR textuels n'exigent pas que leurs bases de connaissances soient dans un format particulier, mais visent plutôt à trouver une réponse à une question en analysant des documents en format texte, tels que les articles de journaux, les documents, les pages Web, les manuels et les encyclopédies. Les SQR textuels répondent à la question par des unités de texte, par exemple, des phrases dans la collection de documents, et, au sein de ces unités, identifient l'élément que la question demande.

4.1. Architecture des SQR textuel

Dans la recherche d'information (RI) et le traitement du langage naturel (TLN), un SQR est la tâche de fournir automatiquement une réponse à une question posée par un humain en langage naturel. Les QR en tant que tâches peuvent être divisées en trois sous-tâches distinctes (Sutcliffe et al. 2013), qui sont l'analyse de questions, l'extraction de documents et l'extraction de réponses (voir la Figure 18). La plupart des SQR suivent ces trois sous-tâches. Cependant, ils peuvent différer dans la façon dont ils mettent en œuvre chaque sous-tâche. D'autres modules peuvent être ajoutés dans les sous-tâches de QR pipeline, par exemple, l'analyse de document, l'extension de requête et la validation de réponse.

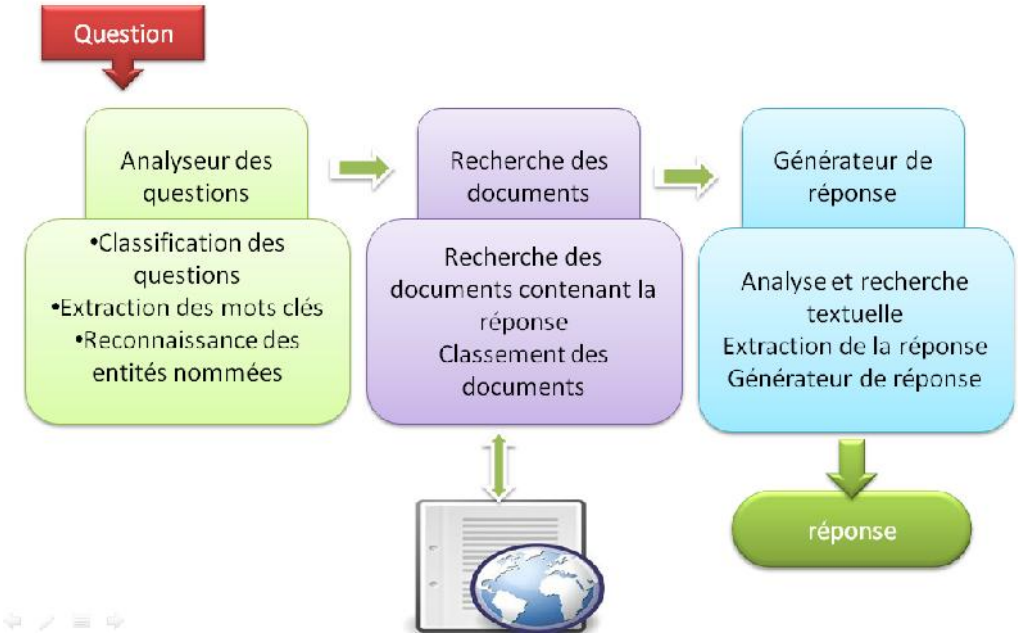

Figure 16: Architecture des SQR pour les données textuelles

Les techniques de traitement du langage naturel sont utilisées pour interfacer le SQR du côté de l'utilisateur qui pose de nombreux types de questions. En particulier, les questions Factoid sont celles qui concernent principalement l'Entité Nommée (NE), en utilisant par exemple les mots : Quand, Où, Combien, Qui et Quoi, qui interrogent respectivement la date/l'heure, le lieu, la personne et l'organisation. Le deuxième type est les questions sur la définition du terme ou du concept. Les questions qui utilisent les mots «Pourquoi» ou «Comment» sont un autre type difficile à répondre, et il y a très peu ou pas de tentatives pour répondre à ce type de questions.

4.1.1. L'analyse de la question

La fonction principale de la composante d'analyse de la question est de comprendre le but de la question, c'est-à-dire, le type d'information que la question demande. Pour identifier le type d'une question, la question est analysée de plusieurs façons. Tout d'abord, le questionnaire reçoit une classe ou un nombre de classes. Le Tableau 5 montre un certain nombre de classes de questions.

Classe de question	Exemples de pattern	Description
Agent	Qui	Nom ou description d'une entité animée provoquant une action. Ex. : Qui a remporté l'Oscar du meilleur acteur en 1970?
Alias	Qu'est-ce que	Nom alternatif pour une entité Ex. : Qu'est-ce que "la peur de la foudre est appelé" ?
Capitale	Quelle est la capitale	Capitale d'un état ou d'un pays. Ex. : Quelle est la capitale du Kentucky?

Date	Quand	Date d'un événement. Ex. : Quand est-ce que l'histoire de ROMEO et JULIET a eu lieu?
Date de naissance	Quand est né	Date de naissance d'une personne. Ex. : Quand est né le roi Louis XIV?
Date de décès	Quand	Date de décès d'une personne. Ex. : Quand Einstein est-il mort?
Exemple d'abréviation	Que représente	La pleine signification d'une abréviation. Ex. : Que représente le NASDAQ?
Location	Où	Emplacement d'une entité ou d'un événement. Ex. : Où est-ce que Chikh Bouamama a grandi?
Objet	Quel est	Une chose identique à la description. Ex. : Quel est le numéro atomique de l'uranium?
What-np	Quel	Une instance du NP correspondant à la description. Ex. : Quel collège Allen a-t-il fréquenté?

Tableau 5: Modèles d'échantillons pour la classification des questions et les types de questions

L'attribution de classes de questions peut être accomplie de diverses manières. L'un des moyens les plus simples, mais tout à fait efficace, consiste à appliquer la correspondance de modèle à la question pour identifier son type. Le tableau 5 répertorie certains modèles utilisés pour classer les questions. La classification est sensible à l'ordre dans lequel les modèles sont appliqués.

Comme alternative, il existe des moyens beaucoup plus sophistiqués de classification des questions. (Suzuki et al. 2002) et (Zhang and Lee 2002) utilisent des SVM (machines à vecteurs de support), une approche d'apprentissage automatique. (Hovy et al. 2001) analysent complètement les questions et ensuite appliquent un grand nombre de règles à l'arbre d'analyse pour classer les questions. (Xu et al. 2014) utilisent des modèles linguistiques pour la classification des questions.

En plus de classer la question, le composant d'analyse de question doit formuler la requête qui est posée au composant de recherche. Pour ce faire, chaque mot est d'abord normalisé à sa racine morphologique. Typiquement, ceci est fait en utilisant un stemmer basé sur des règles, tel que le stemmer de (Porter 1980), ou en recherchant la racine morphologique dans un dictionnaire lisible par machine. Les mots morphologiquement normalisés sont utilisés pour poser la requête au moteur de recherche.

4.1.2. Recherche de Documents

La fonction du composant de recherche de documents n'est pas de trouver des réponses réelles à la question, mais d'identifier les documents susceptibles de contenir une réponse. Ce processus de présélection de documents est également connu sous le nom de pré-extraction. Parce que la réponse à la question est beaucoup plus spécifique que dans la recherche

traditionnelle, de nombreux systèmes utilisent un système de recherche booléen, qui offre plus d'options pour formuler une requête, ou une extraction basée sur les passages textuels qui souligne le fait que les réponses sont normalement exprimées très localement dans un document. L'utilisation d'une approche de recherche de passage au lieu d'une approche de recherche de document complète présente l'avantage supplémentaire de renvoyer des extraits de texte courts au lieu de documents complets, plus faciles à traiter par les composants ultérieurs du SQR.

La recherche de document a une longue tradition et de nombreuses approches ont été développées au fil des années, résultant en des méthodes sophistiquées pour calculer la similarité entre un document et une requête (Salton and Buckley 1988; Zobel and Moffat 2006).

4.1.3. Analyse des documents candidats

L'analyse des documents candidats consiste à fouiller l'information retournée par l'étape de sélection des documents pour identifier des phrases ou des morceaux de phrases correspondante au type d'information recherchée. Les méthodes d'analyse des documents sont les mêmes que celles utilisées en extraction d'information, généralement c'est l'identification des entités nommées et des relations sémantiques. L'identification des entités nommées est utilisée pour identifier le type sémantique de l'information contenue dans les documents candidats. L'identification des relations sémantiques sert à interpréter les liens entre les entités.

La tâche minimale à exécuter lors de cette étape, pour qu'un système fonctionne bien, est l'identification des entités nommées, c'est-à-dire, extraire et classifier les noms de lieux, de personnes, de compagnies, les adresses, les numéros de téléphone, les liens URL et les mesures. D'autres classes et sous-classes peuvent être ajoutées, celles-ci n'étant que les plus communes. Les autres traitements pouvant être exécutés sur les documents sont sensiblement identiques à ceux qui peuvent être appliqués lors du prétraitement de la collection, soit : le découpage en phrase, l'annotation des parties du discours (POS), le «chunking», etc. Lorsque la phrase contient une entité correspondant au type sémantique de l'information recherchée, la phrase est ajoutée à la liste des candidats à considérer pour l'identification de la réponse (Bélanger 2006).

4.1.4. Extraction de la réponse

L'extraction de la réponse est réalisée à partir d'une représentation de la question et des segments de textes retournés par l'analyse des documents candidats. La réponse retournée est rarement unique ; une liste de réponses ordonnées selon une mesure de confiance est généralement plus appropriée. La sélection de la réponse est réalisée de plusieurs manières différentes dans les systèmes question-réponse, selon la représentation de la question, la sélection des candidats et les ressources utilisées par le système.

Une manière de sélectionner les réponses consiste à contraindre les candidats à répondre à des critères de sélection. Le premier critère utilisé est de comparer le type ou la catégorie sémantique de la réponse attendue. Les réponses candidates sont ensuite comparées entre elles pour identifier les réponses les plus fréquentes. La comparaison des fréquences est habituellement effectuée sur le sous-ensemble de documents candidats sélectionnés (Bélanger 2006), mais certains systèmes utilisent le corpus complet pour calculer la fréquence

d'apparition de la réponse avec les termes de la question. Le Web est aussi utilisé pour confirmer la sélection de la réponse en utilisant la fréquence d'apparition de la réponse. La comparaison des réponses selon leur fréquence fonctionne à condition d'admettre l'hypothèse qu'un énoncé plus fréquent est celui qui correspond le plus fidèlement à la réponse. Dans le problème que nous considérons, cette hypothèse ne peut pas être exploitée parce que nous travaillons sur une petite collection de documents spécifiques au domaine, contrairement à la question-réponse traditionnelle qui exploite de gros corpus de documents.

4.2. Travaux sur les systèmes question-réponse

La plupart des travaux en cours sur les SQR ont été repris en grande partie par la conférence TREC Text Retrieval (parrainée par l'American National Institute, le NIST et la DARPA) et par la piste question-réponse multilingue de CLEF.

Comme l'a souligné (Hirschman and Gaizauskas 2001), les SQR pour le texte impliquent essentiellement deux étapes : dans la première étape, l'entité à trouver par la question est définie sémantiquement. Dans la seconde, l'entité de réponse est enrichie par des contraintes supplémentaires.

La liste ci-dessous des travaux de recherche, donne un aperçu concis sur les SQR les plus populaires pour le texte dans la littérature :

LASSO fonctionne en quatre étapes : (1) introduire la question, (2) prédire la réponse, (3) identifier le focus de la question et (4) donner le mot-clé pertinent dans la question et non dans la réponse (Moldovan et al. 1999).

Dans **FALCON**, la reconnaissance des entités nommées est utilisée pour mapper les catégories sémantiques des réponses. Après cette étape, la catégorie de la question est identifiée et elle est mappée vers la taxonomie des réponses (Harabagiu et al. 2000).

Dans **DIMAP**, le document est analysé et converti en triplets (triplets de relation sémantique). Ces triplets sont stockés dans une forme structurée, créant une base de données triplets afin d'être utilisée pour répondre à la question. Les triplets relationnels sémantiques sont extraits à l'aide de techniques sémantiques (Litkowski 2001).

POWER ANSWER (Moldovan et al. 2004) développé au LCC (Language Computer Corporation) cherche des réponses dans une large collection de textes en combinant des sources d'information syntaxiques, sémantiques, lexicales et de connaissances de mots. Ce système comprend trois parties principales : le traitement des questions, la recherche des documents et l'extraction des réponses.

PALANTIR (Harabagiu et al. 2005) a été conçu avec deux objectifs principaux : être une plate-forme de test pour les systèmes question-réponse et être un système convivial (dialogue). Le système PLANTIR utilise plusieurs techniques d'extraction appliquées pour : (1) la détection des collocations, (2) la reconnaissance du type de réponse attendue, (3) l'indexation de la collection de documents basée sur un très grand ensemble de classes d'entités nommées, (4) le classement des réponses, basé sur plusieurs stratégies.

StoQA utilise les techniques du TAL, de reconnaissance des entités nommées (REN), les listes de mots vides et les étiqueteurs de parties du discours pour extraire les informations de

la question de l'utilisateur. Cette phrase sera utilisée comme entrée dans le moteur de recherche travaillant sur le corpus de documents pour trouver le document correspondant contenant la réponse exacte (Stoyanchev et al. 2008).

Mulder est un SQR pour les questions factuelles. La requête utilisateur est étendue à plusieurs requêtes envoyées au moteur de recherche google. Un traitement linguistique utilisant WordNet est effectué pour classer les requêtes, puis un module de formulation convertit chaque requête en un ensemble de mots-clés (Kwok et al. 2001).

QALC fournit des réponses à des questions factuelles anglaises en se basant sur l'analyse syntaxique et sémantique, en utilisant un traitement en langage naturel. Le système QALC utilise six modules : analyse de questions NL, extraction des termes, moteur de recherche, indexation automatique, reconnaissance d'entités nommées et appariement de phrases de questions (Ferret et al. 1999).

QRISTAL utilise massivement les techniques TAL. C'est un système multilingue basé sur les techniques TAL : analyse syntaxique, désambiguïsation sémantique, analyse conceptuelle et thématique et reconnaissance des entités nommées (Laurent et al. 2006).

Web QA utilise la technique de mappage de modèle pour définir le type de la question, et la technique de regroupement pour extraire plusieurs blocs de réponses (Parthasarathy and Chen 2007).

5. Travaux sur les SQR Arabe

Dans le domaine des systèmes question-réponse pour la langue arabe, la situation est moins brillante. Les recherches dans ce domaine sont lentes et donnent des résultats limités pour toutes les sous-tâches du SQR en raison du manque de ressources et d'outils en TAL arabe (Bouziane et al. 2017). Dans ce qui suit, on cite quelques travaux sur les SQR arabes.

Dans (Ahmed and Babu 2016), les auteurs ont proposé un analyseur de questions pour les systèmes question-réponse arabes en utilisant Stanford POS Tagger & Parser pour la langue arabe. Les différents modules sont : (1) Reconnaissance d'entité nommée, (2) Tokenization et (3) Extraction de la bonne réponse.

Al-Bayan (Abdelnasser et al. 2014) est un système question-réponse pour le Saint Coran. L'utilisateur pose des questions en langage naturel arabe à propos du Coran. Le système récupère d'abord les versets les plus pertinents du Coran. Deuxièmement, il extrait le passage qui contient la réponse à partir de deux sources : le Coran et ses livres d'interprétation (Tafseer).

AQuASys (Bekhti and Al-Harbi 2013) est conçu pour répondre à des questions arabes basées sur des faits. Il est composé de trois modules : un module d'analyse de questions, un module de filtrage de phrases et un module d'extraction de réponses.

Dans (Ahmed et al. 2017), la source d'information du système est un corpus donné ou des pages Web. Il utilise un classificateur de machines à vecteur de support « support vector machine » (SVM) supervisé pour la classification des questions et la sélection des réponses, afin de générer la réponse exacte pour une question donnée en langage naturel arabe.

AR2SPARQL (AlAgha and Abu-Taha) est une interface arabe en langage naturel pour le Web sémantique, qui utilise l'analyse linguistique et sémantique pour convertir la question arabe en triplets RDF, qui sont ensuite associés à des triplets d'ontologie pour récupérer une réponse.

AQAS (Mohammed et al. 1993) extrait des réponses à partir de données structurées. C'est le premier système de ce type pour la langue arabe. Le domaine de connaissance des rayonnements est présenté en utilisant la technique des modèles (frames). Il n'y a pas d'évaluation publiée sur AQAS.

QARAB (Hammo et al. 2004) est un système question-réponse non connecté (non basé sur le Web) destiné uniquement à des questions factuelles. Aucun autre type de question n'est pris en charge. Il utilise des techniques de recherche d'information et du TAL pour extraire les réponses d'une collection de textes des journaux arabes.

Dans (Al-Shawakfa 2016), les auteurs utilisent des règles de marquage et des modèles de questions pour analyser et comprendre une question arabe dans un environnement question-réponse.

ArabicQA (Benajiba et al. 2007) est un système question-réponse en langue arabe basé sur un module de récupération des passages, un module de reconnaissance des entités nommées et un module d'extraction des réponses pour les textes arabes.

Dans (Shawar 2011), l'auteur décrit un moyen d'accéder à un corpus de question-réponse Web arabe en utilisant un chatbot (initiative de chatbot open source ALICE). ALICE est l'entité informatique linguistique artificielle. Le système utilise un ensemble simple (mais volumineux) de règles de correspondance de modèle, et convertit un corpus de texte au format de modèle de chatbot AIML.

DefArabQA (Trigui et al. 2017), les auteurs utilisent un schéma lexical pour définir des questions afin d'extraire les infoboxes des l'articles Wikipédia, pour générer des réponses coopératives pour les questions de définition. Cette approche peut être intégrée dans tous les systèmes question-réponse.

Le tableau suivant (Tableau 7) résume les travaux selon quatre critères: (1) "Source de données" indique le nom et le type des données analysées, (2) les "Techniques d'analyse de questions" correspondent aux techniques d'analyse utilisées par le système considéré, (3) Les "Techniques de classification des questions" correspondent aux techniques de classification utilisées et (4) les "performances", qui donnent des résultats expérimentaux.

Système	**Source de données (Texte/document web/ données liées)**	**Techniques d'analyse des questions**	**Techniques de classification des questions**	**Performance du système**
Al bayan (Abdelnasser, et al., 2014)	Coran et livres d'interprétation (non structuré)	Techniques NLP	Classifieur SVM	Évaluation d'experts résultats: 0.73%

	texte.			
AQuASys (Bekhti and Al-Harbi 2013)	Documents (non structurés) texte.	Structure de Question définie et techniques TAL	Question définie Types et formes	Précision:66,25% Rappel: 97,5%
AQAS (Mohammed et al. 1993)	Bases de connaissances du domaine des maladies radiologiques (structurées) données liées.	parseur: analyse morphologique	Non	Non mentionné
QARAB (Hammo et al. 2002), (Hammo et al. 2004)	Texte de journal arabe Al-Raya (non structuré) texte.	Techniques TAL	Utilisant un ensemble de types de questions connues	Précision: 97,3% Rappel: 97,3%
WAHEED (Ahmed et al. 2017)	Web de documents (non structuré)	Techniques Statistique	Classifieur SVM Techniques TAL	MMR (Mean Reciprocal Rank : 65%
AR2SPARQL (AlAgha and Abu-Taha 2015)	Ontologie (structurée) données liées	Techniques TAL	Non	Précision: 85,24% Rappel: 61,61% F-Mesure: 71,5%
AL-SHAWAKFA (Al-Shawakfa 2016),	Corpus de documents (non structuré) texte.	Techniques PNL	Type de question défini	Précision: 78,15% Rappel: 97% F-mesure: 86,56%
ArabicQA (Benajiba et al. 2007)	Ensemble de documents (non structuré) texte.	Techniques TAL	Type de question défini	non mentionné
SHAWAR (Shawar 2011)	Corpus de texte (non structuré)	Règles de concordance	Non	Rappel: 93%.

	texte.	modèle-modèle		
DefArabicQA (Trigui et al. 2017)	Article Wikipedia (semi-structuré) document web.	Motif lexical	Non	Précision: 63%.

Tableau 6: Quelques fonctionnalités et techniques des SQR Arabe.

6. Conclusion

Dans ce chapitre, nous avons présenté un état de l'art sur les systèmes question-réponse, qui sont des systèmes capables de répondre à des questions exprimées en langage naturel. La première génération a connu le développement des interfaces du langage naturel vers les BDD. La génération suivante des systèmes question-réponse pour les données textuelles a exploité le potentiel des données non-structurées. Cette génération a été largement étudiée et utilisée dans le Web actuel. L'évolution du Web 2.0 vers le Web sémantique a montré la nécessité de développement d'une nouvelle génération de SQR, à savoir les systèmes question-réponse pour les données liées (SQRDL). Les SQRDL interrogent les données liées pour répondre aux questions des utilisateurs.

Pour notre étude, nous nous sommes intéressés à l'exploitation des données liées plus particulièrement à un système question réponse pour la langue arabe. La nécessité de crée ce type de système pour les données liée permet a l'utilisateur d'exploité le potentiel du web sémantique en utilisant la langue naturelle arabe.

Chapitre 5 : Un système question-réponse pour les données liées arabes

1. Introduction

Malgré les recherches considérables sur les systèmes question-réponse pour le Web sémantique anglais, le développement de ces systèmes pour la langue Arabe est à ses débuts. En raison des limites dans le TAL arabe (Zaghouani 2017) et dans le Web sémantique arabe, il existe peu de travaux dans le domaine des SQR pour le Web sémantique et l'ontologie arabes (Bouziane et al. 2017). Le web sémantique a comme perspective d'être exploitable par la machine mais aussi accessible a l'humain via un langage naturel. Motivés par ce défi, nous présentons dans cette thèse un système question-réponse pour les données liées arabes.

L'objectif principal de notre système est de fournir des réponses exactes aux questions des utilisateurs exprimées en langage naturel arabe, posées par les utilisateurs du grand publique, qui ne connaissent pas la structure des données et le langage de requête compliqué.

Le reste du chapitre est organisé comme suit. La section 2 met en évidence le défi du développement des systèmes question-réponse pour le Web sémantique arabe. La section 3 passe en revue les travaux existants. La section 4 présente l'architecture de notre système proposé. Un exemple illustratif est présenté dans la section 5. La section 6 évalue la méthode d'extraction des ressources, l'algorithme de suppression des mots vides et la précision globale du système . Enfin, nous résumons le document et mettons en évidence les futures orientations du travail dans la section 7.

2. Motivation et défi

Dans ce chapitre, nous présentons un système question-réponse pour Web sémantique arabe. Le développement d'un tel système pour des ressources pauvres en TAL arabe (Ezzeldin and Shaheen 2012) et en Web sémantique arabe est un défi difficile et a long terme. Nous allons discuter dans ce qui suit les difficultés rencontrées pour mettre en évidence le développement des systèmes question-réponse pour le Web sémantique arabe.

La langue arabe est une collection de plusieurs variantes parmi lesquelles la langue écrite officielle des médias, de la culture et de l'éducation dans le monde arabe. Les autres variantes sont des dialectes parlés informels qui sont les moyens de communication de la vie quotidienne. Bien sûr, le langage existe dans un continuum naturel, à la fois historique et géographique.

Dans le contexte informatique, il existe des ressources et des outils limités pour le traitement de la langue arabe par rapport à la langue anglaise. Ce manque d'outils provoque un manque de prise en charge de la langue arabe dans les technologies du Web sémantique. Actuellement, le Web sémantique arabe est très loin des meilleures performances par rapport au Web sémantique anglais.

Une autre composante importante du Web sémantique est l'ontologie arabe (Boudabous et al. 2013), qui serait la base de la création du Web sémantique arabe. Récemment, plusieurs travaux ont porté sur le Web sémantique pour la langue arabe (Boudabous et al. 2013; AL-Feel 2015). Le Web sémantique arabe a été présenté dans la section 4 du troisième chapitre.

.

La plupart des travaux sur les systèmes question-réponse arabes traitent des données non-structurées. Actuellement, avec le développement de la technologie des données liées, les nouveaux systèmes doivent interagir avec les données liées plutôt qu'avec le Web de documents. Notre système proposé tente de relever ces défis en proposant une nouvelle architecture qui aide à répondre aux questions en arabe en les traduisant en requêtes SPARQL. Dans la section suivante, nous décrivons l'architecture de notre système proposé.

3. Architecture du système

Nous proposons un système question-réponse qui transforme les questions arabes en requêtes SPARQL et fournit ensuite une réponse exacte tirée d'une base de connaissances basée sur une ontologie arabe. Notre système proposé accepte les questions simples de facto en arabe qui peuvent avoir l'un des formats suivants (من هو \ man. Huw ~ a \ qui est, من هي\ man. est, ما هي \ maA hiy ~ a \ ce qui est, متى \ mataý \ quand, مما \ mim ~ aA \ de quoi, أين \ Âay.na \ where, كيف \ kay.fa \ comment, أي في \ fi ay ~ u \ dans quoi, كم \ kam. \ combien, بأي \ biÂay ~ i \ dans quoi…).

Tout d'abord, à travers son interface, le système reçoit la question en NL, la traite et produit finalement une réponse après avoir formulé une requête SPARQL pouvant être exécutée sur le Web sémantique arabe, basé sur une ontologie arabe. Le processus de transformation est composé de trois modules consécutifs. Chacun est composé de plusieurs étapes décrites dans la Figure 22. Il commence par le traitement des questions. Ensuite, la reconnaissance des prédicats est faite. Enfin, la requête SPARQL est formulée.

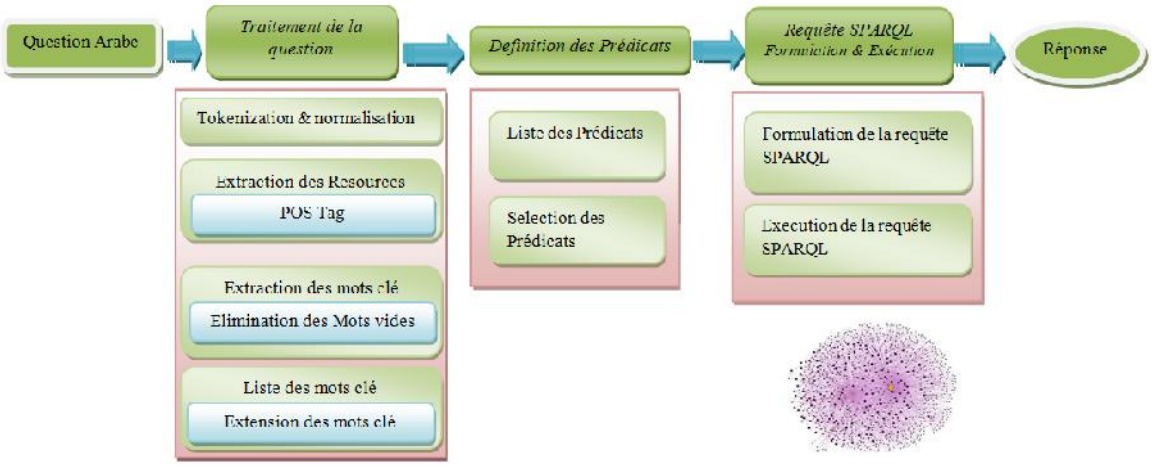

Figure 17: Architecture du système

Dans les sous-sections suivantes, nous expliquons les modules mentionnés ci-dessus (Figure22).

3.1.Traitement des questions

C'est un module important et crucial. Il permet d'analyser la question en entrée exprimée en langage naturel, afin d'avoir ses composants. Ce module a un impact important sur la précision et les performances de tout système question-réponse. Dans notre système proposé, la question d'entrée est traitée et analysée linguistiquement par le module de traitement des

questions, qui comprend quatre étapes : Tokenization et Normalisation, Extraction de ressources, Extraction de mots-clés et Obtention de la liste de mots-clés. Par conséquent, il fournit la liste des mots-clés et les ressources (entités nommées).

a. Tokenization et Normalisation

L'étape commune du TAL est la Tokenization, qui désigne la segmentation du texte en langage naturel (question dans notre cas) en unités de base consécutives individuelles. Une étape de normalisation des mots est nécessaire pour réduire les fautes d'orthographe. Ces erreurs apparaissent parce que la lettre arabe peut être écrite dans des styles différents. La correction des fautes d'orthographe les plus courantes implique la normalisation des caractères Arabe Alif 'ٱ' et Ya 'ي' (Habash 2010). Nous utilisons les outils de MADAMIRA pour l'étape de Tokenization et de Normalisation (Pasha et al. 2014).

Exemple: la tokenization et la normalisation de la question : ما هي عاصمة الجزائر ؟ \maA hiy çaASimah ÂljazaAŷir[1] ?\what is the capital of Algeria ? donne " ما, هي ,عاصمة, الجزائر ". Elle normalise le Alf 'ٱ'et Taa 'ة' dans les mots de lq question.

b. Extraction des ressources

L'entité nommée est très importante dans la plupart des systèmes question-réponse pour les données structurées ou non-structurées (Benajiba et al. 2009). À ce jour, à notre connaissance, aucun système de reconnaissance des entités nommées en arabe n'est disponible gratuitement. Donc, pour combler cette lacune, nous avons mis en place notre propre reconnaissance des entités nommées NER. Dans notre système, l'entité nommée dans la question d'entrée est la ressource cible à explorer dans l'ontologie pour obtenir la bonne réponse.

Ainsi, le processus d'extraction de ressources consiste à extraire la dernière expression nominale (PN) de la question arabe de l'arbre d'analyse syntaxique de la question considérée. Nous utilisons le Tagger Part-of-Speech de Stanford (pos-tagg) tableau 7 présente les différentes labels utilisées par stanford (pos-tagg) qui est une implémentation Java conçue pour fournir une description simple des relations grammaticales dans une phrase (De Marneffe and Manning 2008).

Exemple: la dernière expression nominale (PN) de la question ما هي عاصمة الجزائر؟\maA hiy çaASimaħu AljazaAŷir ?\ est la PN الجزائر \AljazaAŷir\Algeria\Algérie dans la Figure 23.

[1] Notez que dans cet exemple et dans le reste de cette thèse, chaque fois que nous donnons un texte arabe et pour qu'il soit lisible, nous le suivrons d'une translittération HSB Habash, N. Y., A. Soudi and T. Buckwalter (2007). On Arabic Transliteration. Arabic computational morphology: Knowledge-based and Empirical Methods. A. Soudi, A. v. d. Bosch and G. Neumann, Springer. **38:** 15-22. et sa traduction anglaise ou française.

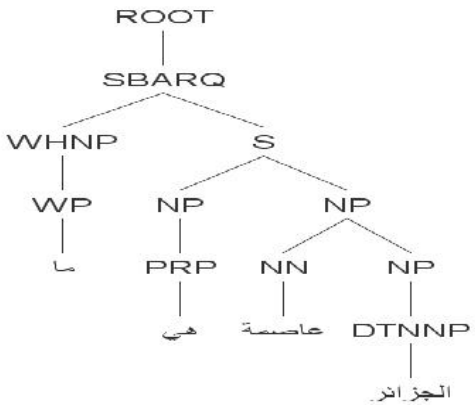

Figure 18: L'arbre sytaxique d'une question arabe, donné par Stanford POS Tagger.

N°	label	Description (anglais)
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle

31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Tableau 7:Liste alphabétique des balises de partie du discours utilisées par le projet Penn Treebank et stanford (pos-tagg)

c. Extraction des Mots-clés

Les mots-clés de la question d'entrée sont utilisés pour générer le prédicat du triplet <Subject, Predicate, Object> et pour formuler la requête SPARQL finale.
Pour extraire les mots-clés de la question d'entrée, le processus se déroule en deux étapes : la première étape consiste à supprimer les mots vides. Les mots vides sont les mots bruyants qui apparaissent fréquemment dans les questions de la langue arabe, tels que les prépositions, les conjonctions et les mots interrogatifs. Nous proposons un automate à états finis (AEF) qui reconnaît les mots vides à supprimer dans notre question arabe d'entrée. La technique des automates à états finis (Figure 24) peut accélérer le processus de suppression des mots vides (Al-Shalabi et al. 2004).

Pour extraire les mots-clés, la deuxième étape consiste à supprimer les mots déjà reconnus lors de l'étape d'extraction des ressources. Les mots-clés sont des mots qui ne sont pas supprimés.

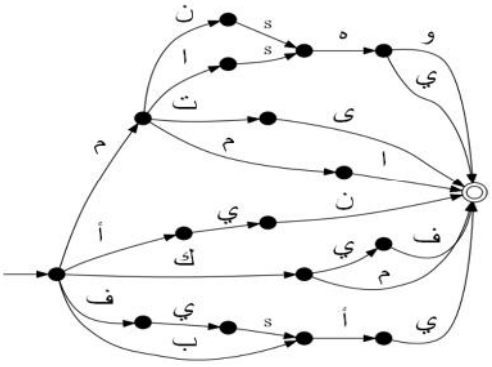

Figure 19: Automate à états finis pour la reconnaissance des mots vides

Exemple: Dans la phrase ما هي عاصمة الجزائر \ maA hiy çaASimaħu AljazaAir \ quelle est la capitale de l'Algérie, l'enlèvement des mots vides donne ما هي \ maA hiy \ quelle est. La ressource est alors الجزائر\ AljazaAŷir \ Algérie. Le reste عاصمة \ çaASimaħu \ capitale est le mot-clé de la question en langage naturel.

d. Extension des mots-clés

Les mots-clés extraits à l'étape précédente et le prédicat utilisé dans l'ontologie peuvent avoir une morphologie différente, mais la même signification. Pour combler cette lacune, nous utilisons WordNet arabe (Rodríguez et al. 2008) (Regragui et al. 2016) pour trouver des

synonymes afin d'étendre la liste des mots-clés. Ces synonymes sont morphologiquement normalisés par le processus de normalisation utilisé dans la première étape de ce module. Une liste de mots-clés est construite pour augmenter la probabilité de définir le prédicat exact dans le module suivant.

3.2. Détermination des prédicats

Une fois que la ressource a été déterminée par le module de traitement des questions, le système utilise une simple requête SPARQL pour obtenir toutes les propriétés de notre ressource.

Question 1: ما هي عاصمة الجزائر؟\maA hiy çaASimaħu AljazaAŷir*what is the capital of Algeria ?*

Question 2: متى ولد هواري بومدين؟ \matý wulida huwaAriy buwmad.yan*What is the birthday of Houari Boumedienne ?*

Les resources pour les questions 1 et 2 sont respectivement الجزائر\AljazaAŷir*Algérie* and هواري بومدين\huwaAriy buwmad.yan*Houari Boumedienne.* Maintenant, nous devons construire une requête SPARQL pour récupérer les propriétés de la ressource à partir de l'ontologie:

Select ?p where{ < http://www.arabic-ontology-2#الجزائر > ?p ?o}

Select ?p where{ < http://www.arabic-ontology-2#هواري بومدين > ?p ?o}

Le fichier SPARQL résultant est un document XML valide par rapport au schéma XML. L'élément principal est le SPARQL. A l'intérieur de l'élément SPARQL, il y a deux sous-éléments: *head* et un élément *result*. L'élément *result* contient la séquence complète du résultat de la requête. Il a un élément de liaison enfant. Dans cet élément de liaison, nous trouvons le nom et la valeur de la variable de requête. Nous utilisons le langage XQuery pour explorer le document XML et créer la liste de prédicats.
Maintenant, nous faisons correspondre la liste de mots-clés avec la liste de prédicats. C'est un processus de correspondance lexicale, où le mot commun est sélectionné pour être le prédicat.

3.3. Formulation et exécution de la requête SPARQL

Maintenant, nous sommes en mesure de formuler la requête SPARQL finale avec la ressource et le prédicat pour fournir la réponse exacte de l'ontologie arabe. Pour cela, nous utilisons la requête de template:

Select ? Object where {Resource Property ?Object}

Pour la question ما هي عاصمة الجزائر؟ \ maA hiy çaASimaħu AljazaAŷir, quelle est la capitale de l'Algérie?, la requête SPARQL est:

Select ? Object where {<http://www.arabic-ontology-2#الجزائر> *onto:* عاصمة *?Object}*

Le résultat de la requête SPARQL est un document XML, qui sera analysé par le même processus de l'étape précédente pour extraire la réponse correcte.

3.4 Algorithme du système :

-lire la question.
-Repeat{
tokenization et normalization des mots
}jusqu'à (fin de question)

-Extraction des ressources

- Stanford pos-tagg (question)
- Extraction de la dernière NP (phrase nominale)

-extraction des mots clé

- Elimination des mots vides (automate)
- Elimination de la ressource.

-Liste des mots clé :

- Extension des mots clé.

-Liste des prédicats :

- Exécution de la requête SPARQL : Select ?p Where {ressource, ?p , ?o}
- Construction de la liste des prédicats.

-Sélection des prédicats :

Correspondance (liste des prédicats, liste des mots clé).

-Requête SPARQL final

- Construction et exécution de la requête SPARQL (Select ?réponse Where {ressource, prédicat, ?réponse}

-génération de réponse.

4. Exemple Illustratif

Maintenant, nous montrons un exemple illustratif complet du traitement d'une question en arabe.

Dans cet exemple, nous supposons que la question introduite dans notre système est la suivante:
امن هو مؤلف رياض الصالحين؟ *Man huwa muŵlif riyaAD AlSAliHiyn? \, Qui correspond à la question en français, qui est l'auteur de The Meadows of the Righteous?*

Le module de traitement des questions se déroule en 4 étapes : Tokenization et Normalization, Extraction de ressources, Extraction de mots-clés et Liste de mots-clés. Enfin, il produit ce qui suit ;

Tokenization et normalisation :
Extraction des Ressources : le résultat de l'arbre syntaxique est : (ROOT (SBARQ (WHNP (WP من*)) (S (NP (PRP*ﻱ*)) (NP (NN* مؤلف *) (NP (NNP* رياض *) (DTNNP5NNP* ﺍﻟ *))))) (PUNC ?))). La ressource est* رياض الصالحين
Extraction des mots clés : *le système élimine les stop-words et la ressource, donc il reste comme mot-clé :* مؤلف
Liste des mots clés : *le système étend les mots-clés définis en utilisant WordNet arabe,* مؤلف *\muŵalif*كاتب *\kaAtib\ auteur, écrivain\ writer, author.*

Le module de traitement des questions produit les informations suivantes pour être l'entrée du module suivant.

Resource: رياض الصالحين*\riyaAD AlSaAliHiyn\The Meadows of the Righteous*

List of keywords: مؤلف*\muŵalif\,* كاتب*\kaAtib\ writer, author.*

Reconnaissance des prédicats: ce module génère et exécute la requête SPARQL via l'ontologie pour récupérer tous les prédicats:
*Select ?p where{ < http://www.arabic-ontology-2#*رياض الصالحين *> ?p ?o}*
Le résultat de la requête SPARQL après traitement est la liste des prédicats: اسم، مؤلف، لغة، كتاب، بلد صدور\Ais.m, muŵalif, luγaħ kitaAb, baladu Suduwr*name, author, language of book, country of publication.*

Ensuite, nous sélectionnons le prédicat : مؤلف \ muŵalif\author comme terme commun entre la liste des prédicats et la liste des mots-clés.
Enfin, le système fournit un résultat de document XML en formulant et en exécutant la requête SPARQL:

SELECT ?object WHERE
{<http://www.semanticWeb.org/ghani/ontologies/2017/6/untitled-ontology-2#
رياض_الصالحين *> onto:*مؤلف *?object }*

La réponse est extraite du document XML :
يحيى_بن_شرف_النووي\yaH.yaý_b.nu_šaraf_Alnawawiy\

5. Evaluation

Avant d'évaluer les performances de notre système proposé, il est important de présenter brièvement le jeu de données utilisé, qui est une ontologie arabe que nous avons développée pour cette évaluation. L'ontologie représentée dans la Figure 25 couvre les notions entourant le concept Person. Un extrait de l'ontologie montre les classes d'ontologie (*Person*, *Occupation*, *Country*, et *Place*), les propriétés de l'objet et les propriétés du type de données « the object properties and the data type properties ».

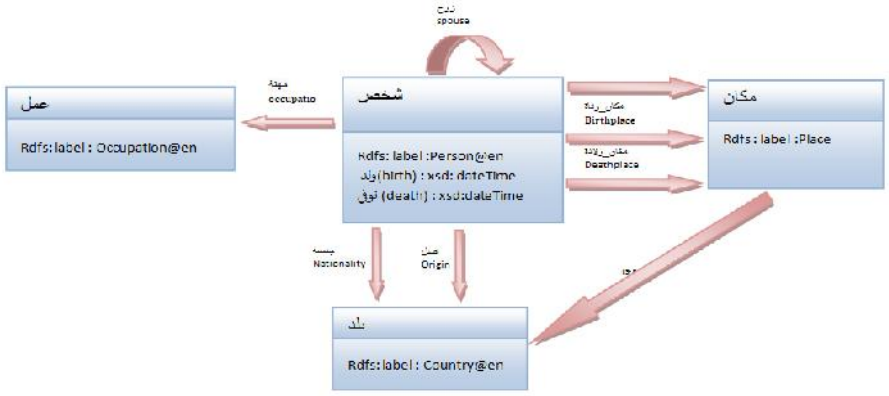

Figure 20: Extrait de l'ontologie Personne

Nous avons effectué plusieurs expériences pour une évaluation complète de notre système. Nous évaluons les performances de la méthode d'extraction des ressources, l'algorithme de suppression des mots vides et la précision globale du système.

Pour évaluer l'extraction des ressources, la suppression des mots vides et le système global, nous utilisons un ensemble de 30 questions simples de type factoid de différents types sur l'ontologie Person.

Cet ensemble de données est collecté à partir de la banque d'essai des tâches de récupération des passages et de réponse aux questions proposées par Yassine Benajiba (Benajiba accessed October 2017).

5.1. Mesures d'évaluation

Pour mesurer les performances de notre algorithme d'extraction de ressources, de suppression de mots vides et de l'ensemble du système, nous avons utilisé les métriques Précision, Rappel et F-mesure.

La précision (P) est définie comme suit :

Extraction des ressources

$$\text{Précision} = \frac{\text{nombre de ressources indentifiées correctement}}{\text{nombre total de ressources génerées par le système}}$$

Suppression des mots vides :

$$\text{Précision} = \frac{\text{nombre de mots vides indentifiés correctement}}{\text{nombre total de mots vides génerés par le système}}$$

Système :

$$\text{Précision} = \frac{\text{nombre de questions correctement traduites}}{\text{nombre total de questions traduites par le système}}$$

Le rappel (R) évalue la couverture, et est définit comme suit :

Processus d'extraction de ressources :

$$\text{Rappel} = \frac{\text{nombre de ressources indentifiées correctement}}{\text{nombre total de ressources}}$$

Suppression des mots vides :

$$\text{Rappel} = \frac{\text{nombre de mots vides indentifiés correctement}}{\text{nombre total de mots vides}}$$

Système :

$$\text{Rappel} = \frac{\text{nombre de questions correctement traduites}}{\text{nombre total de questions introduites}}$$

Enfin, la F-mesure est le compromis entre le rappel et la précision, et elle est calculée en multipliant par 2 le produit Précision et Rappel, divisé par la somme de la précision Précision et le rappel. Mathématiquement, la formule de F-mesure est la suivante: F-Mesure = 2 * (P * R) / (P + R).

5.2. Résultats et discussion

En fait, les évaluations de l'Extraction de ressources, Suppression de mots vides et le système global ont été effectuées à l'aide du jeu de données cité ci-dessous. Les résultats d'évaluation sont décrits dans le tableau 8.

	Extraction des Ressources			**Suppression des mots vides**			**Système**		
Nombre de questions	**Précision**	**Rappel**	**F-mesure**	**Précision**	**Rappel**	**F-mesure**	**Précision**	**Rappel**	**F-mesure**
05	1	1	1	1	1	1	1	1	1
10	0,8	0,88	0,83	0,9	0,95	0,92	0,8	0,8	0,8
15	0,8	0,85	0,82	0,93	0,95	0,94	0,78	0,73	0,75
20	0,75	0,83	0,78	0,9	0,95	0,92	0,77	0,7	0,73

30	0,73	0,7	0,71	0,95	0,95	0,95	0,71	0,66	0,68

Tableau 8: Résultats d'évaluation, exprimés en Précision, Rappel et F-mesure.

5.2.1. Extraction des ressources

Pour le processus d'extraction de ressources, la précision, le rappel et la F-mesure résultants sont respectivement, 0.73, 0.70 et 0.71 sur 1, comme le montre l'histogramme de la Figure 26. La précision et le rappel doivent être améliorés, ce qui reflète la nécessité d'ajouter des règles grammaticales et des listes (gazetteers) dans le processus de reconnaissance. Les cas d'échec peuvent s'expliquer principalement par l'analyse incorrecte de certaines questions en arabe. Les ressources arabes, qui sont des noms arabes, se présentent souvent sous la forme d'expressions nominales, constituées d'un substantif ou d'un adjectif à référence nominale. Dans certains cas, les noms arabes sont plus compliqués et consistent en plusieurs phrases nominales ou phrases verbales.

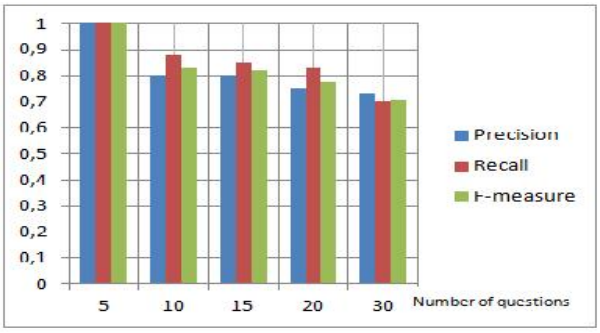

Figure 21: Histogramme d'évaluation de l'extraction des ressources

5.2.2. Suppression des mots vides

La précision de l'algorithme de suppression des mots vides est illustrée dans l'histogramme de la Figure 27. Les résultats d'évaluation (Précision, Rappel et F-mesure) sont élevés (0,95), qui projette une grande précision.

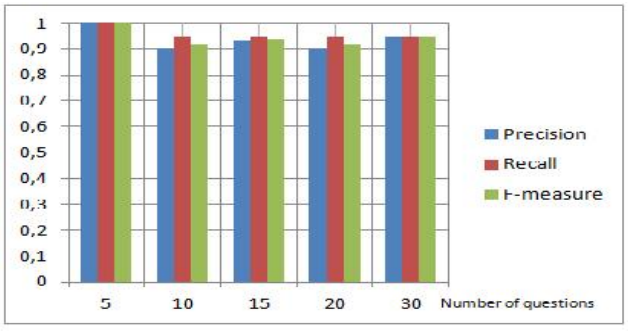

Figure 22: Histogramme d'évaluation de l'élimination des mots vides

5.2.3. Système :

Comme on peut le voir dans l'histogramme de la Figure 28, le système atteint avec succès 0,66 en termes de rappel et 0,71 en termes de précision.

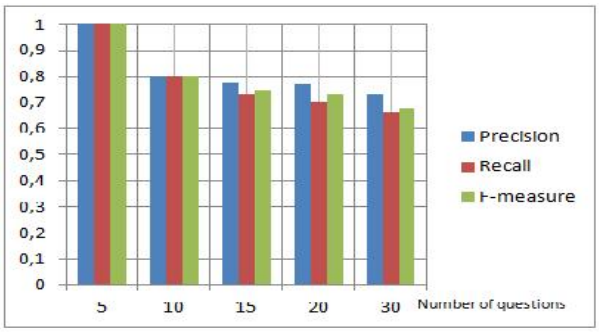

Figure 23: Histogramme d'évaluation du système

Les explications les plus probables des résultats d'échec de notre système sont les suivantes :

- Échec de l'extraction des ressources : les résultats d'évaluation précédente ont montré que le processus d'extraction des ressources a un résultat limité, ce qui affecte négativement la précision du système, comme le montre la Figure 29.

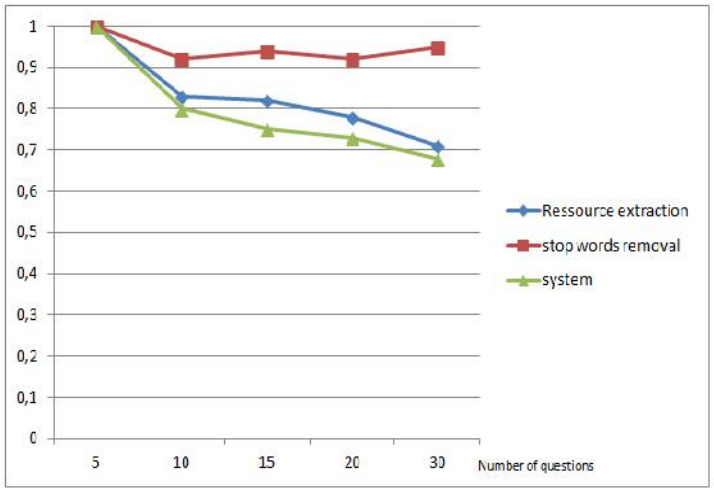

Figure 24: Courbes d'évaluation F-mesure

- Échec des mots-clés : s'est produit lorsque les mots ne pouvaient être associés à aucun prédicat de la ressource appropriée. Les causes les plus fréquentes de ce type d'erreurs peuvent être classées en trois catégories principales :

  - ✓ Échec de la reconnaissance : cette défaillance peut survenir lorsque l'utilisateur pose une question en arabe de manière inhabituelle et non standardisée. N'oubliez pas que notre système traite des questions relatives

à MSA (Modern Standard Arabic). La cause principale de l'échec de la reconnaissance peut être expliquée par l'échec d'extraction des ressources et l'échec de la suppression des mots vides.

- ✓ Échec d'extension : se produit lorsque le système ne génère aucun synonyme adéquat au prédicat à l'aide de WordNet arabe.
- ✓ Échec de correspondance : se produit lorsque la liste des mots-clés ne peut être associée à aucun prédicat. Ainsi, le vocabulaire de l'ontologie doit utiliser les mots les plus fréquents.

5.3. Performances SQR

Dans cette section, nous présentons des statistiques et comparaison sur les SQR basé sur l'ontologie et SQR basé sur le texte en mettant l'accent sur les performances de notre système proposé.

5.4.Systèmes question-réponse basés sur l'ontologie

Pour montrer la performance des SQR basés sur l'ontologie, nous avons examiné les résultats d'évaluation menés dans la littérature, notamment ceux résumés dans le document d'enquête (Lopez et al. 2011) et notre système proposé. Ensuite, nous avons établi l'histogramme de la Figure 19.

Les performances du SQR basé sur l'ontologie sont représentées par le taux de réussite (réponses correctes aux questions) dans le graphe ci-dessous.

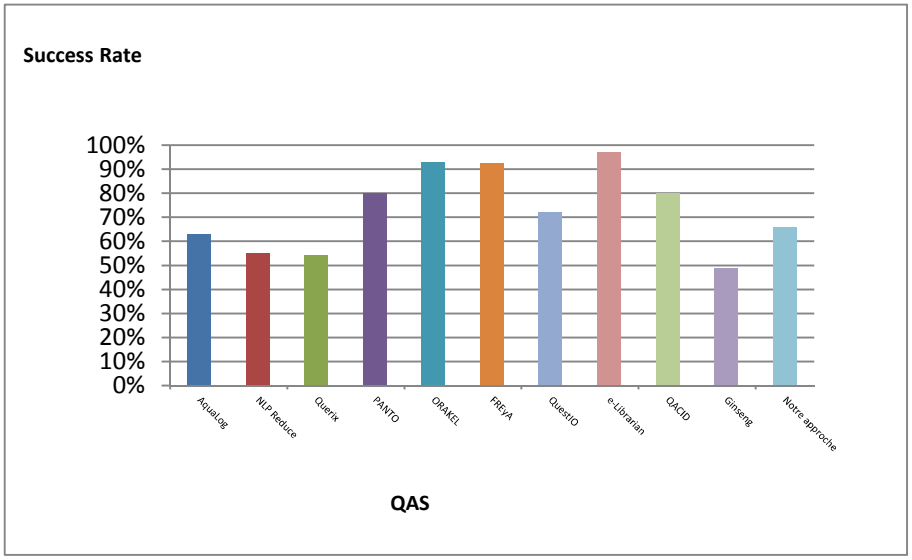

Figure 25: Performance des Systèmes question-réponse basé sur les ontologies

Nous avons constaté que le taux de réussite de ces systèmes QR varie entre 49% et 89%. Tendit que notre système a atteint un taux de réussite de 66%. Ces résultats dépendent de

deux critères : (1) les algorithmes et méthodes de traitement du langage naturel, et (2) le domaine spécifié à interroger.

5.5.SQR basés sur le texte

Pour évaluer les systèmes QR basés sur les textes, nous avons examiné les résultats présentés dans QA4MRE (Question Answering For Machine Reading), qui est la tâche principale du Forum d'évaluation CLEF Cross Langage Evaluation Forum 2013 (Sutcliffe et al. 2013).

QA4MRE lit les documents uniques et identifie les réponses correctes et les réponses NoA à une série de questions, au cours des deux années 2012 et 2013. NoA signifie que le système a décidé de ne pas répondre à la question.

La Figure 20 montre le pourcentage de réponses correctes pour différents types de questions (objectif, méthode, causal, factoid et qui-est-vrai) dans les versions 2012/2013 du challenge QA4MRE.

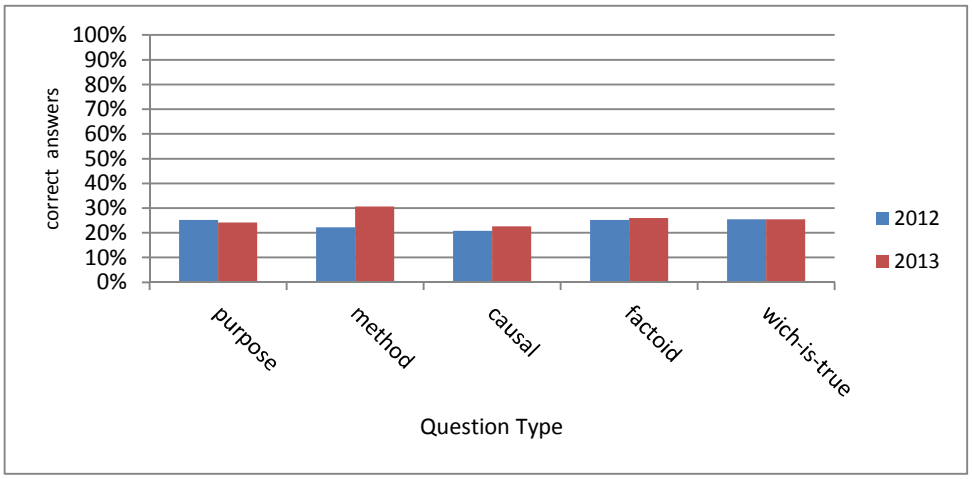

Figure 26 : Pourcentage de réponses correctes pour différents types de questions

En utilisant les mêmes types de questions, la Figure 21 montre le pourcentage de non-réponses en 2012-2013. Un faible nombre de questions sans réponses signifie que le système est plus fiable.

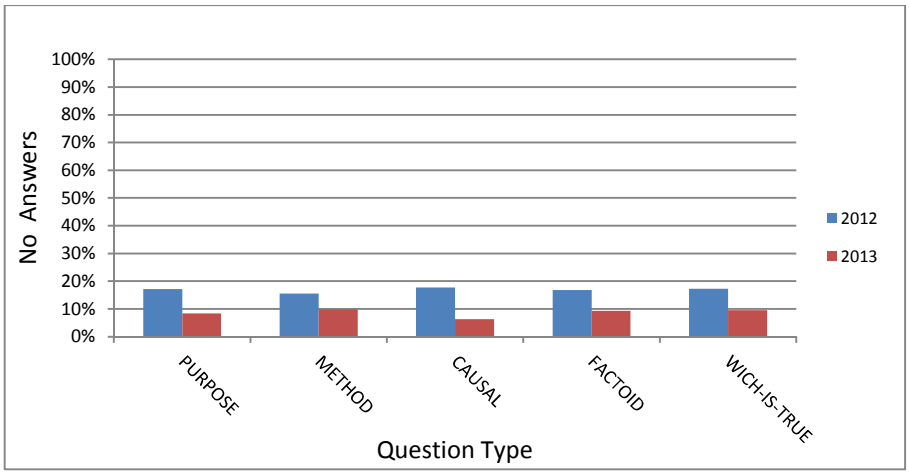

Figure 27: Pourcentage de NoA (pas de réponses) pour différents types de questions

Nous pouvons dire que la fiabilité du SQR augmente proportionnellement au pourcentage de réponses correctes, et est inversement proportionnelle au pourcentage de non-réponses. Ces résultats dépendent de trois critères : le type de questions, la fiabilité de l'algorithme de recherche du document candidat et le module d'extraction des réponses correctes.

Cependant, le fait qu'un SQR ne renvoie pas de réponse ne signifie pas nécessairement qu'il n'est pas capable de trouver une réponse, car parfois, la réponse n'existe pas dans le corpus utilisé. Cela dépend beaucoup de la configuration de l'expérience.

6. Conclusion

Nous avons présenté un système question-réponse qui fournit des réponses aux questions exprimées en langage naturel arabe. Nous pensons que le système proposé permet aux utilisateurs d'explorer le contenu arabe croissant du Web sémantique. Le système utilise des techniques du TAL et du Web sémantique pour traiter la question d'entrée et la transformer en requête SPARQL pour obtenir des réponses du Web sémantique basé sur une ontologie arabe. Tout d'abord, le système applique un processus linguistique à la question saisie afin de fournir la ressource et une liste de mots-clés. Deuxièmement, tous les prédicats de la ressource sont sélectionnés et mis en correspondance avec la liste de mots-clés afin d'identifier le prédicat approprié qui forme le triplet <Sujet, Predicate, Object>. Enfin, notre système formule et exécute une requête SPARQL finale pour obtenir une réponse exacte.

Nous avons discuté les principaux défis du développement de ce type de système pour la langue arabe. Ce qui est précieux pour des études plus approfondies. Sur la base des résultats d'évaluation, on peut conclure que ce domaine de recherche est très prometteur. Cependant, nous prévoyons de résoudre certaines limites que nous avons trouvées dans notre évaluation

en améliorant les modules existants en utilisant des techniques de traitement du langage naturel, ainsi que d'autres outils et ressources.

Étant donné qu'il s'agit de l'un des premiers travaux visant à développé un système question-réponse en arabe par rapport aux données liées, il existe de nombreuses pistes pour étendre notre travail. La première consiste à ajouter de nouveaux modules tels que la catégorisation des questions. Deuxièmement, étendre ce travail à d'autres cas complexes en arabe. Enfin, nous allons appliquer notre approche proposée à un exemple réel du chapitre arabe de DBpedia, qui est l'exemple réel du Web sémantique arabe fourni par la communauté DBpedia.

Chapitre 6 : Conclusion générale

1. Synthèse

La principale motivation de cette thèse était d'étudier les systèmes question-réponse dans le contexte des données liées arabes. Plus précisément, nous souhaitons crées un outil capable d'exploité les données liées arabes et de les rendre accessibles par l'utilisateur simple. Cette outil est appelé communément un système question-réponse sur les données liées arabes. L'objectif des systèmes question-réponse est de répondre correctement aux questions des utilisateurs. Les méthodes traditionnelles pour l'exploitation et la recherche d'information, telles que les moteurs de recherche, les langages de requête, les formulaires et les interfaces graphique, sont couteuses en termes de temps et d'effort, et présentent des limites fonctionnelles. L'évolution de l'informatique et la popularité de l'utilisation du Web nécessitent le développement de nouveaux outils capables de subvenir aux besoins d'information pour les utilisateurs d'une façon plus simple et plus rapide. Actuellement la langue arabe soufre d'un retard considérable dans ce domaine par rapport aux autres langues, telles que l'anglais ou les langues latines en générale. Le Web sémantique arabe est loin d'être dans son potentiel complet. Comme conséquence, il y a une faiblesse dans le développement des systèmes question-réponse Arabe basés sur les données liées.

En informatique, on distingue deux grandes classes de données : les données structurées et les données non-structurées. Les données non-structurées ne sont pas organisées d'une façon prédéfinie et n'ont pas de modèle standard. On cite par exemple les documents textuels et les pages Web HTML. Ce type d'information est fortement présent dans le Web actuel qui est un Web basé sur l'interconnexion des documents et des humains. L'exploitation des données non-structurées se fait principalement à travers les moteurs de recherche. Ces derniers présentent les résultats de recherche sous forme de liens vers des documents. Dans ce cas, le meilleur scénario est que l'utilisateur exprime son besoin sous forme d'une simple question est reçoit une réponse exacte. On parle ici des systèmes question-réponse textuels. Généralement, les SQR textuels suivent une architecture en pipeline de trois modules qui sont l'analyse de questions, l'extraction de documents et l'extraction de réponses. On note que les SQR textuels sont les plus riches en contributions et donnent des résultats pertinents en anglais. L'évolution du Web et les compagnes d'évaluation, telles que TREC, ont fortement soutenu le développement des SQR.

Les données structurées en informatique sont organisées d'une façon prédéfinie et elles sont représentées sous forme de modèles standards. Les bases de données, ontologie OWL, RDF, KBs, sont des exemples de données structurées en informatique. L'exploitation de ces données nécessite la maitrise d'un langage de requête, ce qui est impossible pour la plupart des gens. Une autre façon est d'utiliser les formulaires, mais cette solution reste limitée pour exprimer les besoins des utilisateurs. Pour rendre les données structurées, qui sont des données souvent crédibles et justes, exploitables pour un public non spécialisé, on a besoin de traduire la question de l'utilisateur en requête SQL par exemple. C'est avec cette logique que les premiers SQR, qui sont des interfaces de langage naturel vers les bases de données, ont été réalisés, et c'est avec cette même logique que la nouvelle génération des SQRDL destinés aux données liées et le Web sémantique se développent. L'évaluation des SQRDL présente des résultats pertinents pour l'anglais, le français et d'autres langues européennes.

Pour la langue arabe la situation est loin d'être comparable à l'anglais, et même les langues latines en générale. Le manque des études concrètes et la faiblesse de développement des

outils linguistiques pour la langue arabe, plus l'absence des données d'évaluation standards entravent l'arabisation des technologies informatiques, notamment la recherche dans le domaine des SQR arabes. Le Web sémantique arabe souffre de ce retard technologique de la langue arabe, notamment dans le processus de création des données liées à partir des pages Web. Cette tâche est primordiale pour le Web sémantique qui va étendre le Web actuel vers une nouvelle génération, à savoir le Web 3.0.

Nous avons proposé dans cette thèse un système question-réponse pour les données liées arabe. Le système est basé sur une architecture en pipeline et sur des traitements linguistiques non approfondis. Notre système est capable d'interroger les données liées en arabe en suivant un processus de conversion de questions arabes en SPARQL. Cette version préliminaire de notre système sera le noyau des travaux futures afin d'atteindre notre objectif de rendre le Web sémantique accessible aux utilisateurs ordinaires exprimant leurs besoins sous forme de questions simples en langue arabe. On a constaté que ce domaine est prometteur et prend les avantages des progrès dans les technologies des données liées et le Traitement automatique du langage naturel.

2. Perspectives

La construction d'un système question-réponse utilisable à grande échelle est un projet à long terme. Dans cette thèse, nous avons montré les différents défis auxquels les chercheurs peuvent être confrontés, en particulier pour la langue arabe. Durant la progression de cette thèse, plusieurs points ont surgi et devenu des perspectives de proche, moyen ou long terme, selon la complexité de chaque projet. Les principales orientations que nous proposons sont :

- Perfectionner le module d'Extraction des ressources par l'utilisation des gazetteers, et des règles grammaticaux qui sont en cours de réalisation.

- Eliminer les erreurs de reconnaissance des mots clés, premièrement par l'utilisation des dictionnaires plus adaptés à nos besoins au lieu de WordNet, et deuxièmement par un processus de correspondance sémantique entre les mots clés de la question et les prédicats des triplets.

- Ajouter d'autres modules, tels que la classification des questions en utilisant une nouvelle approche basée sur les ontologies et les structures des phrases arabes.

Une évaluation crédible nécessite la création d'une base de test standard suivant les conférences d'évaluation, telles que QALD-9 qui est la version 2018 de la compétition intitulée « The 4th International Workshop on Natural Language Interfaces for Web of Data (NLIWoD) & 9th Question Answering over Linked Data challenge ». Les questions seront disponibles en 3 à 9 langues différentes (anglais, espagnol, allemand, italien, français, néerlandais, roumain, hindi et farsi), avec éventuellement deux langues supplémentaires (la Corée et le portugais brésilien). Ces questions sont générales, factuelles et à domaine ouvert. Malheureusement, avec l'absence de la langue arabe depuis la création de cette compétition, notre objectif est d'adapter ces bases de test pour la langue arabe.

Conclusion générale

La mise en marche d'un tel système nécessite un cas du monde réel de données liées. En Algérie, Wikipedia représente 62,10%[2] de toutes les références provenant des moteurs de recherche. Le projet DBpedia est basé sur une ontologie universelle provenant des données des articles et infobox de Wikipedia. Un système question-réponse en arabe sur Dbpedia sera notre objectif ultime durant les prochaines années.

[2] Données du mois d'Aout 2018 avec une mise à jour quotidienne. Selon ALEXIA.

Bibliographie

Abdelnasser, H., R. Mohamed, M. Ragab, A. Mohamed, B. Farouk, N. El-Makky and M. Torki (2014). Al-Bayan: An Arabic Question Answering System for the Holy Quran. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), Doha, Qatar, Association for Computational Linguistics.

Ahmed, W. and A. P. Babu (2016). "Question Analysis for Arabic Question Answering Systems." International Journal on Natural Language Computing (IJNLC) **5**(6).

Ahmed, W., A. Pv and A. P. Babu (2017). "Web-Based Arabic Question Answering System using Machine Learning Approach." International Journal of Advanced Research in Computer Science **8**(1): 40-45.

AL-Feel, H. (2015). The Roadmap for the Arabic chapter of DBpedia. MATHEMATICAL and COMPUTATIONAL METHODS in ELECTRICAL ENGINEERING, Proceedings of the 14th International Conference on Telecommunications and Informatics (TELE-INFO '15), Sliema, Malta.

Al-Khalifa, H. S., M. M. Al-Yahya, A. Bahanshal and I. Al-Odah (2009). SemQ: A proposed framework for representing semantic opposition in the Holy Quran using Semantic Web technologies. Current Trends in Information Technology (CTIT), 2009 International Conference on the, IEEE.

Al-Shalabi, R., G. Kanaan, J. M. Jaam, A. Hasnah and E. Hilat (2004). Stop-word removal algorithm for Arabic language. Proceedings of International Conference on Information and Communication Technologies: From Theory to Applications. Damascus, Syria, IEEE**:** 545.

Al-Shawakfa, E. (2016). "ARule-BASED APPROACH TO UNDERSTAND QUESTIONS IN ARABIC QUESTION ANSWERING." Jordanian Journal of Computers and Information Technology **2**(3): 210-231.

Al-Zoghby, A. M., A. S. E. Ahmed and T. T. Hamza (2013). "Arabic semantic web applications: a survey." Journal of Emerging Technologies in Web Intelligence **5**(1): 52-69.

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

Androutsopoulos, I., G. Ritchie and P. Thanisch (1993). "Masque/sql {An E cient and Portable Natural Language Query Interface for Relational Databases." Database technical paper, Department of AI, University of Edinburgh.

Atwell, E. (2015). "A review of morphosyntactic analysers and tag-sets for Arabic corpus linguistics."

Auer, S. and J. Lehmann (2007). What have innsbruck and leipzig in common? extracting semantics from wiki content. European Semantic Web Conference, Springer.

Bas, v. d. B. (2018). "State of digital. The Arabic Web: Numbers and Facts, General statistics." Retrieved September, 2018, from http://www.stateofdigital.com/the-arabic-web/.

Bekhti, S. and M. Al-Harbi (2013). AQuASys: A Question-Answering System For Arabic. 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) Morioka City, Iwate, Japan, WSEAS Press.

Bélanger, L. (2006). "Architecture question-réponse pour l'automatisation des services d'information."

Belkredim, F. Z. and F. Meziane (2008). "DEAR-ONTO: a derivational Arabic ontology based on verbs." International Journal of Computer Processing of Languages **21**(03): 279-291.

Benajiba, Y. (accessed October 2017). Test-Bed for Passage Retrieval (PR) and Question Answering (QUA) tasks. Y. Benajiba.

Benajiba, Y., M. Diab and P. Rosso (2009). "Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition." International Arab Journal of Information Technology (IAJIT) **6**(5).

Benajiba, Y., P. Rosso and A. Lyhyaoui (2007). Implementation of the ArabiQA question answering system's components. Proc. Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morroco, April.

Berners-Lee, T., Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer and D. Sheets (2006). Tabulator: Exploring and analyzing linked data on the semantic web. Proceedings of the 3rd international semantic web user interaction workshop, Citeseer.

Bernstein, A., E. Kaufmann and C. Kaiser (2005). Querying the semantic web with ginseng: A guided input natural language search engine. 15th Workshop on Information Technologies and Systems, Las Vegas, NV, Citeseer.

Beseiso, M., A. R. Ahmad and R. Ismail (2010). "A Survey of Arabic language Support in Semantic web." International Journal of Computer Applications **9**(1): 35-40.

Beseiso, M., A. R. Ahmad and R. Ismail (2011). An Arabic language framework for semantic web. Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on, IEEE.

Binot, J., L. Debille, D. Sedlock and B. Vandecapelle (1991). "Natural language interfaces: a new philosophy." SunExpert Magazine **2**(1): 67-73.

Bizer, C., T. Heath and T. Berners-Lee (2011). Linked data: The story so far. Semantic services, interoperability and web applications: emerging concepts, IGI Global**:** 205-227.

Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann (2009). "DBpedia-A crystallization point for the Web of Data." Web Semantics: Science, Services and Agents on the World Wide Web **7**(3): 154-165.

Borras, J. (2004). "International technical standards for e-government." Electronic journal of e-government **2**(2): 75-80.

Boudabous, M. M., L. H. Belguith and F. Sadat (2013). "Exploiting the Arabic Wikipedia for semi-automatic construction of a lexical ontology." International Journal of Metadata, Semantics and Ontologies **8**(3): 245-253.

Boudlal, A., A. Lakhouaja, A. Mazroui, A. Meziane, M. Bebah and M. Shoul (2010). Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts. International Arab conference on information technology, Benghazi Libya.

Bouziane, A., D. Bouchiha, N. Doumi and M. Malki (2015). "Question answering systems: survey and trends." Procedia Computer Science **73**: 366-375.

Bouziane, A., D. Bouchiha, N. Doumi and M. Malki (2017). "Question answering systems: the story till the Arabic linked data." International Journal of Artificial Intelligence and Soft Computing (IJAISC) **6**(1): 24-42.

Bouziane, A., D. Bouchiha, N. Doumi and M. Malki (2018). "TOWARD AN ARABIC QUESTION ANSWERING SYSTEM OVER LINKED DATA." Jordanian Journal of Computers and Information Technology (JJCIT). **Vol. 04**(No. 02).

Buckwalter, T. (2004a). Buckwalter Arabic Morphological Analyzer Version 2.0. catalog number LDC2004L02, LDC.

Buckwalter, T. (2004b). Issues in Arabic orthography and morphology analysis. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Association for Computational Linguistics.

Burger, J., C. Cardie, V. Chaudhri, R. Gaizauskas, S. Harabagiu, D. Israel, C. Jacquemin, C.-Y. Lin, S. Maiorano and G. Miller (2001). Issues, tasks and program structures to roadmap research in question & answering (Q&A). Document Understanding Conferences Roadmapping Documents.

Cabrio, E., J. Cojan, A. P. Aprosio, B. Magnini, A. Lavelli and F. Gandon (2012). QAKiS: an open domain QA system based on relational patterns. International Semantic Web Conference, ISWC 2012.

Cimiano, P., P. Haase and J. Heizmann (2007). Porting natural language interfaces between domains: an experimental user study with the orakel system. Proceedings of the 12th international conference on Intelligent user interfaces, ACM.

Cloud, I. M. (217). key marketing trends for 2017 and ideas for exceeding customer expectations.

Codd, E. (1974). Seven Steps to RENDEZVOUS with the Casual User. J. Kimbie and K. Koffeman, editors, Data Base Management, North-Holland Publishers.

Damljanovic, D., M. Agatonovic and H. Cunningham (2011). FREyA: An interactive way of querying Linked Data using natural language. Extended Semantic Web Conference, Springer.

Davies, J., D. Fensel and F. Van Harmelen (2003). Towards the semantic web: ontology-driven knowledge management, John Wiley & Sons.

DBpedia. (2018). "Towards a Public Data Infrastructure for a Large, Multilingual, Semantic Knowledge Graph." Retrieved 06/09/, 2018, from https://wiki.dbpedia.org/services-resources/ontology.

De Marneffe, M. C. and C. D. Manning (2008). Stanford typed dependencies manual Stanford University.**:** 338-345.

Dietz, L. and B. Gamari. (2017). "trec car "trec-car.cs.unh.edu"." Retrieved 24 November 2017, 2017, from trec-car.cs.unh.edu.

Dima, C. (2014). Answering Natural Language Questions with Intui3. CLEF (Working Notes).

Doumi, N. (2017). extraction des connaissance a partir du texte docteur en science, UNIVERSITE DJILLALI LIABES SIDI BEL ABBES

Embregts, H., V. Milea and F. Frasincar (2014). Metafrastes: A news ontology-based information querying using natural language processing. The 8th International Conference on Knowledge Management in Organizations, Springer.

Emine, Y., K. Evangelos, C. Nick, B. Peter, C. Ben and M. Rishabh (2017). "Tasks Track."

Engström, G. (2008). "Internationalisation and Localisation Problems in the Chinese and Arabic Scripts." Uppsala University, Sweden.

Eugene, A., A. Ben Abacha, D. Harman, E. Nyberg and Y. Pinter (2017). "Trec LiveQA ".

Ezzeldin, A. M. and M. Shaheen (2012). A survey of Arabic question answering: challenges, tasks, approaches, tools, and future trends. Proceedings of The 13th International Arab Conference on Information Technology (ACIT 2012).

Fan, J., A. Kalyanpur, D. C. Gondek and D. A. Ferrucci (2012). "Automatic knowledge extraction from documents." IBM Journal of Research and Development **56**(3.4): 5: 1-5: 10.

Fensel, D. (2000). "Ontologies: A silver bullet for knowledge management and electronic-commerce (2000)." Berlin: Spring-Verlag **143**.

Ferrández, O., R. Izquierdo, S. Ferrández and J. L. Vicedo (2009). "Addressing ontology-based question answering with collections of user queries." Information Processing & Management **45**(2): 175-188.

Ferré, S. (2013). squall2sparql: a Translator from Controlled English to Full SPARQL 1.1. Work. Multilingual question answering over linked data (QALD-3).

Ferret, O., B. Grau, G. Illouz, C. Jacquemin and N. Masson (1999). QALC - the Question-Answering program of the Language and Cognition group at LIMSI-CNRS. TREC-8. Columbia, NIST special publication.

Grace Hui, Y., S. Ian, L. Jiyun and T. Zhiwen. (2017). "TREC Dynamic Domain Track 2017." from http://trec-dd.org/index.html.

Green, B. F., A. K. Wolf, C. Chomsky and K. Laughery (1961). BASEBALL: An automatic question answering. Proceedings Western Joint Computer Conference, McGraw-Hill.

Gruber, T. R. (1993). "A translation approach to portable ontology specifications." Knowledge acquisition **5**(2): 199-220.

Guo, R. and F. Ren (2009). Towards the relationship between Semantic Web and NLP. International Conference on Natural Language Processing and Knowledge Engineering. Dalian**:** 1-8.

Habash, N. (2010). Introduction to Arabic natural language processing, Morgan & Claypool.

Habash, N. and R. Roth (2008). Identification of Naturally Occurring Numerical Expressions in Arabic. LREC.

Habash, N. Y., A. Soudi and T. Buckwalter (2007). On Arabic Transliteration. Arabic computational morphology: Knowledge-based and Empirical Methods. A. Soudi, A. v. d. Bosch and G. Neumann, Springer. **38:** 15-22.

Hammo, B., H. Abu-Salem and S. Lytinen (2002). QARAB: A question answering system to support the Arabic language. SEMITIC '02 Proceedings of the ACL-02 workshop on Computational approaches to semitic languages Philadelphia, Pennsylvania, Association for Computational Linguistics Stroudsburg, PA, USA.

Hammo, B., S. Abuleil, S. Lytinen and M. Evens (2004). "Experimenting with a question answering system for the Arabic language." Computers and the Humanities **38**(4): 397-415.

Harabagiu, D. S., A. Harabagiu, D. Moldovan, C. Clark, M. Bowden, A. Hickl and P. Wang (2005). Employing Two Question Answering Systems in TREC 2005. Proceedings of the Fourteenth Text REtrieval Conference.

Harabagiu, S. M., S. J. Maiorano and M. A. Paşca (2003). "Open-domain textual question answering techniques." Natural Language Engineering **9**(3): 231-267.

Harabagiu, S. M., D. I. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. C. Bunescu, R. Girju, V. Rus and P. Morarescu (2000). FALCON: Boosting Knowledge for Answer Engines. TREC.

Harris, S., A. Seaborne and E. Prud'hommeaux (2013). "SPARQL 1.1 query language." W3C Recommendation **21**(10).

Hazman, M., S. R. El-Beltagy and A. Rafea (2009). "Ontology learning from domain specific web documents." International Journal of Metadata, Semantics and Ontologies **4**(1-2): 24-33.

He, S., Y. Zhang, K. Liu and J. Zhao (2014). CASIA@ V2: A MLN-based Question Answering System over Linked Data. CLEF (Working Notes).

Heflin, J. (2004). "OWL Web Ontology Language-Use Cases and Requirements." W3C Recommendation **10**: 12.

Hendrix, G., E. Sacerdoti, D. Sagalowicz and J. Slocum (1978). " Developing a Natural Language Interface to Complex Data." ACM Transactions on Database Systems **3**(2): 105–147.

Hirschman, L. and R. Gaizauskas (2001). "Natural language question answering: the view from here." Natural Language Engineering **7**(4): 275-300.

Horrocks, I. (2008). "Ontologies and the semantic web." Communications of the ACM **51**(12): 58-67.

Hovy, E., U. Hermjakob and C.-Y. Lin (2001). The use of external knowledge in factoid QA. TREC.

Huang, R. and L. Zou (2013). Natural language question answering over RDF data. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, ACM.

Ittycheriah, A., M. Franz, W.-J. Zhu, A. Ratnaparkhi and R. J. Mammone (2000). IBM's Statistical Question Answering System. TREC.

Katz, B. (1993). START, the world's first Web-based question answering system. MIT Computer Science and Artificial Intelligence Laboratory, InfoLab Group

Kaufmann, E. and A. Bernstein (2007). How useful are natural language interfaces to the semantic web for casual end-users? The Semantic Web, Springer**:** 281-294.

Kirk, R., W. Hersh, D. Demner-Fushman , E. Voorhees, A. Lazar and S. Pant (2017). "TREC Precision Medicine / Clinical Decision Support Track."

Kocaleva, M., D. Stojanov, I. Stojanovic and Z. Zdravev (2016). "Pattern Recognition and Natural Language Processing: State of the Art." TEM Journal **5**(2): 236-240.

Kwok, C., O. Etzioni and D. Weld (2001). Scaling question answering to the Web. Proceeding. of the 10th International Conference on World Wide Web, Hong Kong, China, ACM.

Laurent, D., P. Séguéla and S. Nègre (2006). Cross lingual question answering using qristal for clef 2006. Workshop of the Cross-Language Evaluation Forum for European Languages, Springer.

Lehnert, W. G. (1977). A conceptual theory of question answering. Proceedings of the fifth International Joint Conference on Artificial Intelligence, Cambridge, Massachusetts,, San Francisco, California: Morgan Kaufmann Publishers. .

Li, Y., H. Yang and H. Jagadish (2006). Constructing a generic natural language interface for an XML database. EDBT, Springer.

Linckels, S. and C. Meinel (2005). A Simple Solution for an Intelligent Librarian System. Proceedings of the IADIS International Conference of Applied Computing 2005 (IADIS AC2005), Lisbon, Portugal.

Litkowski, K. C. (2001). "Syntactic clues and lexical resources in question-answering." NIST SPECIAL PUBLICATION SP(249): 157-166.

Lopez, V., M. Fernández, E. Motta and N. Stieler (2012). "Poweraqua: Supporting users in querying and exploring the semantic web." Semantic web **3**(3): 249-265.

Lopez, V., V. Uren, E. Motta and M. Pasin (2007). "AquaLog: An ontology-driven question answering system for organizational semantic intranets." Web Semantics: Science, Services and Agents on the World Wide Web **5**(2): 72-105.

Lopez, V., V. Uren, M. Sabou and E. Motta (2011). "Is question answering fit for the Semantic Web? A survey " Semantic web **2**(2): 125–155.

Maamouri, M., A. Bies, T. Buckwalter and W. Mekki (2004). The penn arabic treebank: Building a large-scale annotated arabic corpus. NEMLAR conference on Arabic language resources and tools, Cairo.

Microsoft. (2018). "Arabic Tools kit (ATK)." 2018, from https://www.microsoft.com/en-us/research/project/arabic-toolkit-service-atks/?from=http%3A%2F%2Fresearch.microsoft.com%2Fen-us%2Fprojects%2Fatks%2F.

MINIWATTS, M. G. (2018). "Internet World Users By Language : Top 10 Languages." Retrieved september, 2018, from http://www.internetworldstats.com/stats7.htm.

Mishra, A. and S. K. Jain (2016). "A survey on question answering systems with classification." Journal of King Saud University-Computer and Information Sciences **28**(3): 345-361.

Moawad, I. F., M. Abdeen and M. M. Aref (2010). Ontology-based architecture for an arabic semantic search engine. The Tenth Conference. On Language Engineering Organized by Egyptian Society of Language Engineering (ESOLEC'2010).

Mohammed, F. A., N. Khaled and H. M. Harb (1993). "A knowledge based Arabic question answering system (AQAS)." ACM SIGART Bulletin **4**(4): 21-30.

Moldovan, D., S. Harabagiu, M. Pasca, R. Mihalcea, R. Girju, R. Goodrum and V. Rus (2000). The structure and performance of an open-domain question answering system. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics.

Moldovan, D., S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju and V. Rus (1999). Lasso: A tool for surfing the answer net. TREC.

Moldovan, D., S. Harabagui, C. Clark, M. Bowden, J. Lehmann and J. Williams (2004). Experiments and analysis of lcc's two qa systems over trec 2004. Proceedings of The Thirteenth Text Retrieval Conference (TREC 2004), Gaithersburg, USA.

Moldovan, D., M. Paşca, S. Harabagiu and M. Surdeanu (2003). "Performance issues and error analysis in an open-domain question answering system." ACM Transactions on Information Systems (TOIS) **21**(2): 133-154.

Neelu Nihalani, S. S., Mahesh Motwani (2011). "Natural language Interface for Database: A Brief review " IJCSI International Journal of Computer Science Issues **8**(2).

Ngonga Ngomo, A.-C., L. Bühmann, C. Unger, J. Lehmann and D. Gerber (2013). SPARQL2NL: verbalizing sparql queries. Proceedings of the 22nd International Conference on World Wide Web, ACM.

Park, S., S. Kwon, B. Kim and G. G. Lee (2015). ISOFT at QALD-5: Hybrid Question Answering System over Linked Data and Text Data. CLEF (Working Notes).

Parthasarathy, S. and J. Chen (2007). A web-based question answering system for effective e-learning. Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on, IEEE.

Pasha, A., M. Al-Badrashiny, M. Diab, A. E. Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow and R. M. Roth (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. LREC2014.

Popescu, A.-M., A. Armanasu, O. Etzioni, D. Ko and A. Yates (2004). Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability. Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics.

Porter, M. F. (1980). "An algorithm for suffix stripping." Program **14**(3): 130-137.

Pradel, C., O. Haemmerlé and N. Hernandez (2014). Swip: A Natural Language to SPARQL Interface Implemented with SPARQL. 21st International Conference on Conceptual Structures, ICCS 2014. N. Hernandez, R. Jäschke and M. Croitoru. Iaşi, Romania, Springer International Publishing Switzerland**:** 260-274.

PROMISE. (2017). "CLEF (Conference and Labs of the Evaluation Forum)." Retrieved 29 November 2017, from http://www.clef-initiative.eu/.

Ravichandran, D. and E. Hovy (2002). Learning surface text patterns for a question answering system. Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics.

Regragui, Y., L. Abouenour, F. Krieche, K. Bouzoubaa and P. Rosso (2016). Arabic WordNet: New Content and New Applications. Proceedings of the Eighth Global WordNet Conference.

Resnik, P. (1989). Access to multiple underlying systems in JANUS, BBN SYSTEMS AND TECHNOLOGIES CORP CAMBRIDGE MA.

Robert, W., N. C. David, L. Marc, M. James, M. James and W. Dekai (1988). "THE BERKELEY UNIX CONSULTANT PROJECT " Computational Linguistics **14**(4).

Rodríguez, H., D. Farwell, J. Farreres, M. Bertran, M. Alkhalifa, M. A. Martí, W. Black, S. Elkateb, J. Kirk, P. Vossen and C. Fellbaum (2008). Arabic WordNet: Current State and Future Extensions. Proceedings of the Fourth Global WordNet Conference (GWC 2008), Szeged, Hungary.

Ryu, P.-M., M.-G. Jang and H.-K. Kim (2014). "Open domain question answering using Wikipedia-based knowledge model." Information Processing & Management **50**(5): 683-692.

Salahudeen, T., A. S. Ajibola and A. J. Romoke (2010). "Barihi Adetunji, Ph. D."

Saleh, L. M. B. and H. S. Al-Khalifa (2009). AraTation: an Arabic semantic annotation tool. Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services, ACM.

Salim, J., S. F. M. Hashim and A. Aris (2010). A framework for building multilingual ontologies for Islamic portal. Information Technology (ITSim), 2010 International Symposium in, IEEE.

Salton, G. and C. Buckley (1988). "Term-weighting approaches in automatic text retrieval." Information Processing & Management **24**(5): 513-523.

Scha, R. J. H. (1977). "Philips Question Answering System PHILIQA1." ACM SIGART Bulletin, **61**: 26-27.

Schuth, A. and K. Balog (2017). "TREC OpenSearch – Academic Search Edition."

Sharaf, A., E. Atwell, K. Dukes, M. Sawalha, A. Al-Saif, S. Sharoff, K. Markert, L. Al-Sulaiti, B. Abu Shawar and N. Abbas (2010). Arabic and Quranic computational linguistics projects at the University of Leeds المشاريع الحاسوبية على اللغة العربية والقرآن بجامعة ليدز. Proceedings of the workshop of Increasing Arabic Contents on the Web, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), Leeds.

Shawar, B. A. (2011). "A Chatbot as a natural web Interface to Arabic web QA." International Journal of Emerging Technologies in Learning (iJET) **6**(1): 37-43.

Simmons, R. F. (1965). "Answering English questions by computer: a survey." Communications of the ACM **8**(1): 53-70.

Stoyanchev, S., Y. C. Song and W. Lahti (2008). Exact phrases in information retrieval for question answering. Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering, Association for Computational Linguistics.

Sutcliffe, R., A. Peñas, E. Hovy, P. Forner, Á. Rodrigo, C. Forascu, Y. Benajiba and P. Osenova (2013). Overview of QA4MRE Main Task at CLEF 2013. Working Notes CLEF.

Suzuki, J., Y. Sasaki and E. Maeda (2002). SVM answer selection for open-domain question answering. Proceedings of the 19th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics.

Tablan, V., D. Damljanovic and K. Bontcheva (2008). A natural language query interface to structured information. European Semantic Web Conference, Springer.

Tacchini, E., A. Schultz and C. Bizer (2009). Experiments with wikipedia cross-language data fusion. Workshop on Scripting and Development, Citeseer.

Toutanova, K., D. Klein and C. Manning (2013). "Stanford Core NLP." The Stanford Natural Language Processing Group.

trec-core (2017). Common Core " https://trec-core.github.io/2017/".

Trec. (2017). "TREC Real-Time Summarization Track." from http://trecrts.github.io/.

Trigui, O., L. H. Belguith and P. Rosso (2017). "Arabic Cooperative Answer Generation via Wikipedia Article Infoboxes."

Unger, C. and P. Cimiano (2011). Pythia: Compositional meaning construction for ontology-based question answering on the semantic web. International Conference on Application of Natural Language to Information Systems, Springer.

Voorhees Ellen M and Tice DM (2000). Overview of the TREC-9 Question Answering Track. TREC.

Voorhees, E. M. and D. K. Harman. (2017). "Text REtrieval Conference (TREC)." Retrieved 2 November, 2017.

Voorhees, E. M. and D. M. Tice (1999). The TREC-8 Question Answering Track Evaluation. NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC 8) 83.

Wang, C., M. Xiong, Q. Zhou and Y. Yu (2007). Panto: A portable natural language interface to ontologies. European Semantic Web Conference, Springer.

Warren, D., Pereira, F. (1982). " An efficient and easily adaptable system for interpreting natural language queries " Computational Linguistics **8**(3-4): 110-122.

WINOGRAD, T. (1971). Procedures as a representation for data in a computer program for understanding natural language.

Wong, Y. W. (2005). Learning for semantic parsing using statistical machine translation techniques, Computer Science Department, University of Texas at Austin.

Woods, W., R. Kaplan and B. Webber (1972). The Lunar Sciences Natural Language Information System. Bolt Beranek and Newman Inc., Cambridge, Massachusetts Final Report. **Report No 2378**.

Woods, W. A. (1968). Procedural Semantics for a Question-Answering Machine. Proceedings of the Fall Joint Computer Conference.

Xu, K., S. Zhang, Y. Feng and D. Zhao (2014). Answering Natural Language Questions via Phrasal Semantic Parsing. Natural Language Processing and Chinese Computing, Third CCF Conference, NLPCC 2014. C. Zong, J.-Y. Nie, D. Zhao and Y. Feng. Shenzhen, China, Springer-Verlag Berlin Heidelberg 2014. **496:** 333-344.

Yu, L. (2011). A developer's guide to the semantic Web, Springer Science & Business Media.

Zaghouani, W. (2017). "Critical survey of the freely available Arabic corpora." arXiv preprint arXiv:1702.07835.

Zaidi, S., M. T. Laskri and K. Bechkoum (2005). A cross-language information retrieval based on an Arabic ontology in the legal domain. Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems (SITIS'05).

Zhang, D. and W. S. Lee (2002). Web Based Pattern Mining and Matching Approach to Question Answering. TREC.

Zobel, J. and A. Moffat (2006). "Inverted files for text search engines." ACM computing surveys (CSUR) **38**(2): 6.

Résumé :

L'intérêt croissant pour le traitement automatique de la langue Arabe et pour les recherches sur le Web sémantique a engendré un besoin émergeant de développer des nouveaux Systèmes Question-réponse (SQR). Ces systèmes permettent aux utilisateurs de poser une question en langage naturel et obtenir une réponse exacte. Cependant, la plupart des systèmes question-réponse existants se basaient sur les langues latines et l'anglais. Moins d'efforts ont été consacrés à la langue arabe, qui appartient aux "langues sémitiques". Dans cette thèse, nous abordons le problème de développement d'un système question-réponse, et nous proposons une première version d'un nouveau système question-réponse arabe indépendant du domaine, interrogeant des données liées arabes, et visant à aider les utilisateurs arabes à explorer un Web sémantique qui utilise une ontologie arabe. Nous décrivons avec suffisamment de détail les différents modules de notre système proposé, qui utilise des techniques de traitement automatique du langage naturel et des techniques du Web sémantique pour traiter et répondre de manière linguistique à la question exprimée en langage naturel arabe. Des expériences ont été menées pour évaluer et montrer l'efficacité du système proposé.

Mots clés : Système Question-réponse (SQR), Traitement Automatique de la Langue (TAL), Données Liées, Langue Arabe, Web sémantique, Ontologie, SPARQL.

Abstract:

The increasing interest in Arabic natural language processing and semantic Web research involves an emerging need to the development of new Question Answering Systems (QAS). These systems allow users to ask a question in Arabic natural language and get the relevant answer. However, most existing QA systems focused on English and Latin-based languages. Less effort has been concentrated on the Arabic language, which belongs to "Semitic Languages". In this thesis, we address the issue of developing a question answering system and propose an early version of a new domain-independent Arabic question answering system over linked data, which aims to particularly help Arab users to explore the Arabic Semantic Web based on Arabic ontology. We describe with sufficient details the different modules of our proposed system, which uses Arabic natural language processing and semantic Web techniques to linguistically process and answer Arabic natural language question. Experiments have been carried out to evaluate and show efficiency of the proposed system.

Keywords: Question Answering System (QAS), Natural Language Processing (NLP), Linked Data, Arabic Language, Semantic Web, Ontology, SPARQL.

ملخص : إن الاهتمام المتزايد بالمعالجة الآلية للغة العربية والأبحاث حول الويب الدلالي أدى إلى الحاجة العاجلة لتطوير أنظمة جديدة للإجابة عن الأسئلة. تسمح هذه الأنظمة للمستخدمين بطرح سؤال باللغة العربية والحصول على الإجابة المناسبة. ومع ذلك، فإن معظم أنظمة الإجابة عن الأسئلة الموجودة تركز على اللغة الإنجليزية واللغات اللاتينية. حيث أنه قد بذلت مجهودات أقل حول اللغة العربية التي تنتمي إلى "اللغات السامية". في هذه الأطروحة، نتطرق لمسألة تطوير أنظمة الإجابة عن الأسئلة ونقترح نسخة أولية لنظام جديد مستقل عن أي ميدان للإجابة عن الأسئلة العربية حول البيانات المرتبطة العربية، والذي يهدف بشكل خاص إلى مساعدة المستخدمين العرب على استكشاف الويب الدلالي الذي يعتمد على انطولوجيا عربية. نوضح بالتفصيل الوحدات المختلفة لنظامنا المقترح، الذي يستخدم تقنيات المعالجة الآلية للغة، و كذا تقنيات الويب الدلالي للمعالجة اللغوية و الإجابة عن الأسئلة باللغة العربية. أجريت تجارب لتقييم وتوضيح كفاءة النظام المقترح.

الكلمات المفتاحية: نضام الإجابة عن الأسئلة، المعالجة الآلية للغة العربية، البيانات المرتبطة، اللغة العربية، الويب الدلالي، الانطولوجيا، SPARQL.