REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
CENTRE UNIVERSAIRE DE NAAMA –SALHI AHMED –

DEPARETEMENT MTHEMATIQUE ET INFORMATIUE
FILIERE: Informatique.

Mémoire

En vue de l'obtention du diplôme de Master Académique

Intitulé

Annotation des textes arabes avec Linked Data.

Présenté par :

- **Azzizi Zeyneb.**

Soutenu devant le jury composé de :

Dr ……..

Dr ……..

Dr ……

Année universitaire : 2020 /2021.

Dédicace

Je dédie ce modeste travail :

À mon défunt père d'amour

Papa, tu as laissé un grand vide dans ma vie, sachant qu'il y aura toujours une place pour moi dans mon cœur. Même si tu ne me sembles pas, je ne peux pas te toucher, te voir et je ne sais pas que tu veilleras toujours sur moi, comme toujours. Papa, tu me manques, et j'espère que tu es fier de moi que je sois à toi.

Résumé

Le besoin d’applications capables de gérer intelligemment la surcharge d’information disponible sur le Web a été aggravé par une explosion vertigineuse de nombres de pages non exponentiels. Ce besoin est encore plus capitalistique dans certaines tâches qui nécessitent une manipulation sémantique de documents et de contenus en langage naturel ou par capitalisation humaine dans de fins domaines de spécialité. Les ontologies représentent une voie prometteuse pour résoudre ce problème. Leur construction manuelle s’est avérée trop onéreuse et très peu réutilisable. Les structures semi-automatisées commencent à donner des résultats encourageants, elles sont relativement faciles à développer et sont plus partageable et plus réutilisable. L'ontologie en arabe est pratiquement inexistante, pourtant l'arabe est une langue parlée par plus de 435 millions de personnes dans plus de 22 pays, localisés dans la région MENA (Moyen-Orient et Afrique du Nord). L'arabe est la quatrième langue utilisée sur le Web, 50,3% de la population étant des internautes, ce qui représente 5,3% des utilisateurs mondiaux, donc pour la langue arabe, la situation actuelle est moins brillante; Le contenu arabe sur le Web ne reflète pas l'importance de cette langue. Étant donné que l'arabe est l'une des langues les plus importantes du Web et qu'elle souffre malheureusement d'un manque de ressources, il est donc essentiel de créer des ressources linguistiques pour elle maintenant. **Notre objectif, dans ce mémoire** est de développé une approche linguistique pour annoter la collection de textes arabes en utilisant des données liées, en particulier DBpedia, qui est liée à des données ouvertes (LOD) extraites de Wikipédia. Cette approche utilise des techniques de langage naturel pour mettre en évidence le texte arabe avec les données ouvertes associées. Les résultats de l'évaluation de cette approche sont encourageants, malgré la grande complexité de notre base de connaissances indépendante et les faibles ressources en traitement de la langue arabe naturelle. [1]

Abstract

The need for applications capable of intelligently handling the information overload available on the Web has become urgent in the face of the dizzying explosion in the number of pages which continues to grow exponentially. This need is even more essential in certain tasks which require the manipulation of the content and semantics of natural language documents or in the capitalization of human expertise in areas of fine specialties. Ontologies represent a promising way to meet this challenge. Their manual construction has proven to be too expensive and not very reusable. Semi-automatic construction is starting to give encouraging results, given the relative ease of developing them and being more shareable and reusable. Ontologies in the Arabic language are almost non-existent, yet Arabic is a language spoken by more than 435 million people in more than 22 countries, located in the MENA region (Middle East and North Africa). Arabic is the fourth language used on the Web, 50.3% of the population being Internet users, which represents 5.3% of global users, so for the Arabic language, the current situation is less bright; Arabic content on the web does not reflect the importance of this language. Since Arabic is one of the most important languages on the web and unfortunately suffers from a lack of resources, it is therefore essential to create language resources for it now. Our objective in this thesis is to develop a linguistic approach to annotate the collection of Arabic texts using linked data, in particular DBpedia, which is linked to open data (LOD) extracted from Wikipedia. This approach uses natural language techniques to highlight Arabic text with associated open data. The results of the evaluation of this approach are encouraging, despite the great complexity of our independent knowledge base and the limited resources in natural Arabic language processing.

ماخص

لقد اصبحت الحاجة إلى التطبيقات التي بإمكانها التعامل بذكاء مع الكم المتزايد من المعلومات المتاحة على شبكة الإنترنت عاجلة مع انفجار عدد مذهل من الصفحات الذي يستمر في النمو بشكل كبير في كل يوم. لذلك اصبحت هذه الحاجة أكثر أهمية في بعض المهام التي تتطلب منا التعامل مع المحتوى الدلالي المكتوب باللغة الطبيعية أو تثمين الخبرة البشرية في بعض المجالات ذات التخصصات الدقيقة.

لذلك تعتبر الأنطولوجيات وسيلة واعدة لمواجهة هذا التحدي. وتبين أن البناء اليدوي لهذه الانطولوجيات باهظ الثمن زيادة على ذلك فانه من الناذر إعادة استعمالها.

بدأ الشبه الأوتوماتيكي في عطاء نتائج مشجعة ، نظرا إلى السهولة النسبية في تطويرها و امكانية اعادة استعمالها و تقسيمها.

الأنطولوجيات في اللغة العربية تكاد تكون معدومة ، رغم ان اللغة العربية هي لغة التي يتحدث بها أكثر من 300 مليون شخص عبر أكثر من 22 دولة ، فهي رابع لغة مستخدمة على شبكة الإنترنت، حيث أن 50.3٪ من السكان هم من مستخدمي الإنترنت ، وهو ما يمثل 5.3٪ من مستخدمي العالم ، لذلك بالنسبة للغة العربية ، فإن الوضع الحالي أقل إشراقًا.

هدفنا في هذه الاطروحة هو تطوير نهج لغوي لتوضيح النصوص العربية باستخدام البيانات المرتبطة بالخصوص DBpedia ، المرتبط بالبيانات المفتوحة (LOD) المستخرجة من ويكيبيديا. يستخدم هذا النهج تقنيات اللغة الطبيعية لتسليط الضوء على النص العربي مع البيانات المفتوحة المرتبطة به.

نتائج تقييم هذا النهج مشجعة ، على الرغم من التعقيد الكبير لقاعدة المعرفة المستقلة لدينا والموارد المحدودة في معالجة اللغة العربية الطبيعية.

Remerciements

Je remercie **Allah** de m'avoir donné la santé et la volonté d'entamer et de terminer ce mémoire.

Tout d'abord, ce travail ne serait pas aussi riche et n'aurait pas pu avoir le jour sans l'aide et l'encadrement de Mr **Abdelghani BOUZIANE**, je le remercie pour la qualité de son encadrement exceptionnel, pour sa patience, sa rigueur et sa disponibilité durant notre préparation de ce mémoire.

Je remercie s'adresse également à tous mes professeurs pour leurs générosités et la grande patience et pour son aide pratique et son soutien moral et ses encouragements.

A la lumière de mes jours, la source de mes efforts, la flamme de mon cœur, ma vie et mon bonheur, à toi **Ma mère**.

Aux personnes qui j'ai bien aimé à tous mes frères : **NADJIB**, **ABD ELHEQ**, **MOHAMMED**, **ABD ELRAHIM**, **ABD ELSTTAR** et mes amis : **ASMAA** et **AICHA**, je dédie ce travail dont le grand plaisir leurs revient en premier lieu pour leurs conseils, aides, et encouragements.

Aux personnes qui m'ont toujours aidé et encouragé, qui étaient toujours à mes côtés, et qui m'ont accompagnaient durant mon chemin d'études.

Table des matières

Liste des tables

Liste des figures

1. Introduction

1.1. Problématique

Le traitement automatique de la langue arabe a généré de nombreuses recherches scientifiques au cours des deux dernières décennies. L'arabe TAL ou ANLP pour Arabic Natural Language en anglais, a connu une fièvre chez les scientifiques dans les grands laboratoires de recherche et les grandes universités de l'Université de Stanford et surtout était l'Université de Pennsylvanie aux États-Unis après les événements du 11 septembre 2001. Cet engouement est intensifié davantage avec l'apparition des réseaux sociaux qui peuvent être considérés comme un grand média d'échange entre le grand public sauf que ce grand public utilise pour la communication des langues vernaculaires non standards (les dialectes de l'arabe). En plus des points cités ci-dessus s'ajoute l'intérêt que donnent ALESCO (Arab League Educational, Cultural and Scientific Organization) et tous les pays arabes et particulièrement les pays de Golf à l'enrichissement et l'arabisation du contenu web et à l'informatisation de leurs gouvernements (gouvernement électronique). Un autre facteur à cet engouement, c'est la place qu'occupe la langue arabe dans le classement mondial des langues les plus utilisées sur Internet (la quatrième place devançant le français et l'allemand par cinq positions). Ce classement montre clairement que ce langage se développe rapidement en termes d'utilisateurs sur Internet. Le monde arabe en tant que marché offre une opportunité aux grandes entreprises internationales de rechercher, de développer et d'appliquer cette langue dans leur informatique. Le Web sémantique est un Web de données (W3C, 2019), le type de données que l'on trouve dans les bases de données. La collecte d'ensembles de données interdépendants sur le Web peut également être appelée données liées, qui sont renforcées par des technologies telles que RDF et SPARQL. RDF fournit la base pour la publication et la liaison des données. SPARQL est le langage de requête du Web sémantique. Comme il n'y a pas suffisamment de ressources Web sémantiques arabes disponibles sur le Web, nous nous référons aux étiquettes arabes de DBpedia, qui est la version RDF Linked Data de Wikipédia. [2][3]

1.2. Motivation

Les ontologies en langue arabe sont quasi inexistantes, nous pouvons citer quelques travaux encore en herbe [5] et [6] et au fond de laboratoires dénombrable sur les doigts. **Pour le succès du Web sémantique arabe**, nous devons redoubler nos efforts pour générer des outils et parvenir à la représentation au niveau sémantique de l'information permettant à la machine d'être le citoyen de premier ordre sur le Web. Cependant, les principales limitations de l'annotation sémantique en langue arabe sont les technologies Web de la PNL arabe et de la sémantique arabe. Tous deux souffrent du manque de ressources qui affecte négativement le développement du Web sémantique et des applications basées sur le langage naturel. Dans ce qui suit, nous abordons les difficultés liées à la PNL arabe et côté Web sémantique. [5][6]

La contribution originale de notre travail est de faire plus d'efforts dans ce domaine pour combler cette lacune et permettre à la langue arabe de satisfaire les besoins des utilisateurs arabes sur le Web.

1.3. Plan du mémoire

Ce mémoire est organisé ainsi :

Dans le chapitre 1, on présente la langue arabe et ses principales caractéristiques influant sur son traitement automatique.

Le chapitre 2 est consacré au l'évolution du web traditionnel vers le web sémantique.

Le chapitre 3 est consacré aux ontologies, les méthodologies existantes, les stratégies de construction, les formalismes et les langages utilisés.

Le chapitre 4 est un état de l'art, nous y abordons les méthodes d'extractions de termes et de relations et donne un bilan des outils existants pouvant être adaptés à l'Arabe ainsi que leur performance, on présente la méthodologie proposée pour la notation de texte arabe avec linked data (description du système).

A la fin, nous présentons les perspectives et les conclusions données à ce travail de fin d'étude de master.

Chapitre1 : Traitement automatique de la langue arabe.

1. Introduction

La langue arabe est la langue des populations arabes qui firent leur entrée dans l'histoire depuis 3 millénaires environ et qui occupaient les zones septentrionales de l'Arabie.

La langue arabe est considérée comme la 5éme langue courante utilisée dans le monde et la quatrième langue utilisée sur le Web, 50,3% de la population étant des internautes, ce qui représente 5,3% des utilisateurs mondiaux. [7]

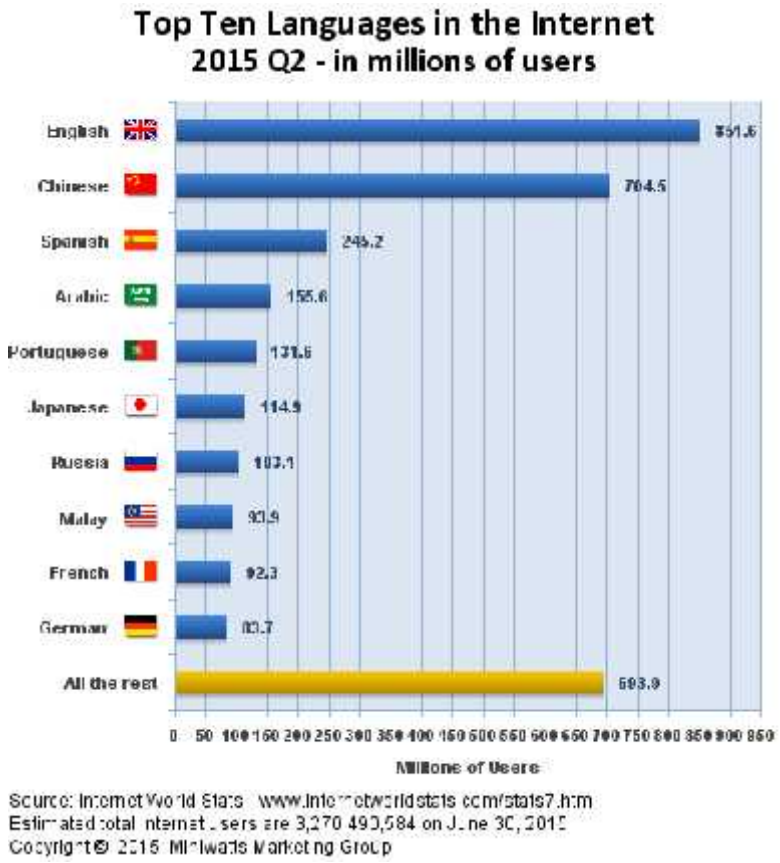


Figure 1 : Les dix langues les plus utilisées sur Internet.

2. Langue arabe :

La langue arabe est l'une des six langues officielles et de travail des Nations Unies. Plus de 435 millions de personnes parlent la langue arabe dans 22 pays arabes localisés dans la région MENA (Moyen-Orient et Afrique du Nord). L'arabe est la quatrième langue

utilisée sur le Web, 50,3% de la population étant des internautes, ce qui représente 5,3% des utilisateurs mondiaux

Le système archaïque d'écriture arabe était consonantique. Chaque lettre de l'alphabet arabe représente une consonne unique depuis les temps anciens. Cependant, la fin du VIIe siècle,

Les diacritiques arabes qui sont des symboles graphiques qui discriminent entre la variété des prononciations des consonnes ont été inventés par "Abou Al-Aswad Al-Du'ali".

Néanmoins, ils sont très souvent éliminés du texte écrit d'aujourd'hui. Lecteurs arabes pouvaient discerner les mots avec la même forme d'écriture par l'intermédiaire de son contexte.

Diviser le texte d'entrée en fractions désirées est généralement la phase initiale dans la plupart des tâches de traitement de texte.

Ces fractions pourraient être des phrases, des chiffres, des mots, des caractères ou toute autre fraction utile. Chaque fraction est appelée un **« Token »** et le processus est appelé **« Tokenization ». [4]**

3. Statut géographique de la langue arabe

L'arabe est une langue parlée par plus de 200 millions de personnes. Elle est langue officielle d'au moins 22 pays :

- Péninsule arabique : l'Arabie saoudite, Bahreïn, les Émirats Arabes Unis, Oman, le Qatar, le Yémen.
- Moyen-Orient : l'Irak, la Jordanie, le Koweït, le Liban, la Palestine, la Syrie.
- Afrique : l'Algérie, l'Égypte, les Comores, Djibouti, la Libye, le Maroc, la Mauritanie, la Tunisie, la Somalie, le Soudan.

C'est aussi la langue de référence pour plus d'un milliard de musulmans. [7]

4. Les caractéristiques et la Complexité de la langue arabe

L'arabe est une langue difficile pour un certain nombre de raisons :

- Composé de 28 lettres (25 consonnes et 3 voyelles longues) (غ ف ق ك ل م ن ه و ي أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع), (vingt-neuf (29) lettres si on n'a pas exclu la hamza (الهمزة), qui se comporte soit comme une lettre à part entière soit comme un diacritique).

- Dans la langue arabe, il n'y a pas de majuscules ou minuscules pour les lettres comme les lettres anglaises et s'écrit de droite à gauche.
- La langue arabe a une morphologie très complexe par rapport à la langue anglaise.
- Les signes diacritiques (الحركات) dans la langue arabe permettent d'exprimer les voyelles brèves et d'apporter différentes modulations aux voyelles longues ainsi qu'aux consonnes.
- L'ambiguïté vocalique des mots, ce qui constitue une grande source d'ambiguïté dans les textes arabes. [1]

5. Les Diacritiques arabes :

Les signes diacritiques (الحركات) dans la langue arabe permettent d'exprimer les voyelles brèves et d'apporter différentes modulations aux voyelles longues ainsi qu'aux consonnes.

Le but de signes diacritiques pour apprendre à les reconnaitre et à les prononcer correctement en contexte, pour distinguer des lettres ambigües et pour faciliter la lecture.

La majeure partie de l'écriture arabe est écrite sans **Harakat**. Cependant, ils sont couramment utilisés dans certains textes religieux qui exigent le strict respect des règles de prononciation telles que Qur'an (القران). Il est fréquent d'ajouter Harakat à hadiths (الحديث), ainsi une autre utilisation dans la littérature pour enfants pour connaitre le sens des mots arabes. **Harakat** sont également utilisés dans les textes ordinaires quand une ambiguïté de la prononciation pourrait se poser.

Les Diacritiques arabes comprennent :

Fatha (ً) (الفتحة), Kasra(ِ) (الكسرة), Damma(ُ)(الضمة), Soukoune(ْ)(السكون) Shadda (ّ) (الشدة)et Tanwin(ٌ) (التنوين).

5.1. Fatha :

Permet la réalisation de la voyelle brève [a]. Il se présente sous la forme d'un accent aigu placé juste au-dessus de la lettre.

5.2. Damma :

Permet la réalisation de la voyelle brève [u].Il se présente sous la forme d'un mini waw (و) placé juste au-dessus de la lettre

5.3. Kasra

Permet la réalisation de la voyelle brève [i]. Il se présente sous la forme d'un accent aigu placé juste en dessous de la lettre

5.4. Soukoune

Les syllabes peuvent être ouvertes ou fermées. C'est-à-dire si la syllabe se termine par une consonne, elle est fermée. Si la syllabe se termine par une voyelle, elle est ouverte. Pour indiquer qu'une syllabe est fermée (à la prononciation), on ajoute simplement un soukoune (petit cercle) au-dessus de la lettre.

5.5. Tanouine

Tanouinefatha ou fathatan : permet la réalisation du son [an]. Il se présente sous la forme d'un double fatha.

Tanouinedamma ou dammatan : permet la réalisation du son [on].Il se présente sous la forme d'un double damma.

Tanouinekasra ou kasratan : permet la réalisation du son [en] ou [an]. Il se présente sous la forme d'un double kasra. [3]

6. Les catégories du mot

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

6.1 Le verbe

Le verbe est une entité qui exprime un sens variant en nombre, en personne et en temps, exemple : **شاهد**; sa conjugaison dépend du temps, du nombre, du genre, de la personne et du mode, il peut donc être exprimé à l'accompli ou l'inaccompli, au singulier, duel ou pluriel, au masculin ou au féminin, au premier, deuxième ou troisième type et être au mode actif ou inactif.

6.2. Le nom

Le nom est un élément qui désigne une chose qui représente une signification indépendante du temps, exemple : **كتاب.**

Il peut être propre, commun ou dérivé d'un verbe. Il s'exprime au singulier, au duel ou au pluriel, au féminin ou au masculin. Il peut être agent, objet, instrument ou lieu.

6.3. La particule

La particule est une entité qui sert à situer les événements par rapport au temps et par rapport à l'espace. Elles peuvent être des conjonctions de coordination « **و/أو/ام...** » ou de subordination «**اذا لان..**». Les particules sont généralement des mots outils, bien que jouant un rôle important dans la cohésion d'une phrase, sont souvent associées à des mots vides qui ne véhiculent pas un sens spécifique à un domaine donné. [1]

7. La structure morphologique d'un mot arabe

La morphologie est une partie importante du traitement du langage :

Une implémentation complète et précise de sujets de morphologie pré-générés et facilite les tâches directement liées telles que le point automatique, étiquetage, classification des extractions de racines, etc.

Ou des applications de haut niveau telles que l'information, la traduction automatique, la synthèse automatique, etc. [1]

7.1. La dérivation

L'Arabe est une langue générative, les noms et les verbes sont dérivés d'une racine, généralement, trilitère. Nous pouvons engendrer jusqu'à 150 mots différents à l'aide de schèmes et ce, à partir d'une même racine. Le tableau1 donne quelques schèmes du mot «**شهد**» **. [8]**

Schème	**شهد**	
فَعَلَ	شَهَدَ	il a témoigné
فَعِلَ	شَهِدَ	Il assisté
فَاعَلَ	شَاهَدَ	Il a regardé
فَاعِلٌ	شَاهِدٌ	Témoin
مَفْعَلٌ	مَشْهَدٌ	Scène
فُوعِلٌ	شُوهِدَ	Il a été vu
فَعَالَةٌ	شَهَادَةٌ	Témoignage, certificat
فَعِيل	شَهِيد	Martyr

Tableau 1: Quelques schèmes du mot "شهد"

7.2. Cas de graphie

L'Arabe s'écrit de droite à gauche avec la particularité que les lettres épousent des formes différentes selon qu'elles soient au début, au milieu ou à la fin du mot, le tableau2 illustre le script de quelques lettres dans les trois cas de graphie.
Cependant, Il faut noter que certaines lettres ne s'attachent pas à celles qui la succèdent comme { **ا ، ذ ،د ،ر ،س ، ص**}. [1][8]

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

Tableau 2: Etat de transcription des lettres arabes

7.3. La voyellation

L'absence de la voyellation est très souvent une grande source d'ambiguïté pour l'analyse morphologique, syntaxique, sémantique et même pragmatique. La majorité des textes écrits, exception faite pour les textes sacrés et quelques ouvrages pédagogiques, sont non voyellés. Cette ambiguïté réside dans le fait que74% des mots qui composent le vocabulaire arabe, acceptent plus d'une voyellation lexicale, et 89,9% des noms qui le constituent acceptent plus d'une voyellation casuelle. La proportion des mots ambigus passe de 90,5% si les comptages portent sur leurs voyellations globales [9]
Prenons l'exemple du mot **شهد** et ses voyellations données par le dictionnaire *almaani*

شَّهْدُ	Miel (cire d'abeille)
شَهَدَ	Informer, affimer, a été présent , a vu
شَهُدَ	A fait une déposition
شُهُّدٌ	Comme رُكَّعٌ
شَهْدٌ	Pluriel de des témoins
شَهْدٌ	Nom propre féminin ,Plante

Tableau 3: Les différentes voyellations du mot "شهد ".

7.4. Flexion des noms/adjectifs

Comparés aux verbes, les morphèmes nominaux sont complexes et hétérogènes. Les noms arabes sont flexibles selon le genre, le nombre et la casse.

7.5. Genre et nombre

L'arabe a deux valeurs de genre : masculin et féminin ; et trois valeurs de nombres : singulier, double et pluriel. Cependant, en ce qui concerne leur forme, l'histoire est compliquée.

- Pluriel irrégulier par exemple **مكتب**\mak.tab*bureau* singulier masculin, **مكاتب**\makaAtib*bureaux* pluriel masculin,
- Féminin irrégulier par exemple **ازرق**\Âaz.raq*bleu* masculin singulier,**زرقاء** \zar.qaA'*bleue* féminin singulier,

7.6. Cas

Les noms arabes fléchissent suivant le cas qui a trois valeurs : nominatif, accusatif et génitif, le cas nominal est réalisé dans la plupart des cas par les signes diacritiques.

7.7. Segmentation

Dans un mot arabe, la base est généralement entourée de propositions et de pronoms qui s'agglutinent à la racine en tant que préfixes, suffixes, infixes, antéfixes ou postfixes, de telle sorte qu'un mot arabe peut résumer à lui seul, toute une phrase exprimée dans une autre langue telle que le Français par exemple, le tableau 4, montre un exemple de segmentation d'un mot arabe.

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

ـه	ـوتـ	لكـ	ـتـ	مـ	تـ	آسـ
			Infixe			
Postifxe	Suffixe	Corps schématique			Préfixe	Antéfixe

Tableau 4: Exemple de segmentation d'un mot arabe.

- Les antéfixes sont des prépositions ou conjonctions (question, futur..) ;
- Les préfixes, infixes et suffixes expriment les traits grammaticaux et indiquent :
  - Cas du nom ;
  - Mode du verbe (actif, passif);
  - Modalités : nombre (singulier, duel, pluriel), genre (Masculin, Féminin), personne (1,2 ou 3 type) ;
- Les postfixes sont des pronoms personnels [8]

8. La richesse de la langue arabe

Selon Ernest Renan, les Arabes se targuent d'avoir 80 appellations de miel, 200 pour les serpents, 500 pour les lions, 1000 pour les chameaux et les épées, et 4400 pour le retour d'idées des malheureux.

Les grammairiens arabes considèrent toutes les racines de la langue comme des verbes, et ils augmentent le nombre de ces racines. [3]

9. La situation de langue arabe sur le web

Consacré au traitement automatique de la langue arabe, mais les différents problèmes posés par la langue et les villes morphologiquement et spécifiques Leurs graphismes ont ralenti le développement d'outils dans ce domaine, donc :

- La situation actuelle est moins lumineuse
- Le contenu de la langue arabe sur le Web ne reflète pas l'importance de cette langue.
- Elle manque malheureusement de ressources. [10]

10. L'importance de la langue arabe

L'importance de la langue arabe est résumée dans ce qui suit :

- La langue arabe est la langue religieuse pour les musulmans du monde entier.
- L'arabe est la langue du Coran.
- C'est la langue maternelle de vingt-trois pays.
- Il y a 300 millions de personnes qui parlent l'arabe comme langue principale.
- C'est l'une des six langues officielles des Nations Unies.
- L'arabe est l'un des langages sémantiques les plus riches qui ont des mots spécifique pour décrire une chose spécifique. [10]

11. Conclusion

Ce chapitre présente les principales particularités de la langue arabe et qui peuvent expliquer la différence au niveau du traitement automatique entre l'Arabe en tant que langue sémitique et les autres langues. Et problèmes se posent lorsqu'il s'agit de la langue arabe naturelle, ce qui pourrait avoir affecté la création de ces techniques et outils de PNL, en particulier sur le Web. La langue arabe est caractérisée par des complexités morphologiques grammaticales et sémantiques, un langage très flexionnel et dérivationnel, l'absence de majuscules, une forte ambiguïté. [4]

.

Chapitre 2 : L’évolution du Web traditionnel vers le web sémantique.

1. Introduction

Le Web actuel est essentiellement syntaxique, dans le sens où la structure du document est bien définie, mais que le contenu est pratiquement impossible d’accès au traitement machine. Seuls les humains peuvent interpréter leur contenu. Le Web Nouvelle Génération – Le Web Sémantique vise à surmonter cette difficulté. Les ressources Web seront plus facilement accessibles par les humains que par les machines, grâce à la sémantique de leur contenu. En particulier, le web sémantique est d’abord une infrastructure d’utilisation des connaissances formalisées en plus des contenus informels du web, même s’il n’y a pas de consensus sur ce que doit être la formalisation. Cette infrastructure doit permettre d'abord de localiser, de définir et de transférer des ressources robustes et saines, et de soutenir l'esprit d'ouverture du Web à la diversité des utilisateurs. Elle doit reposer sur un certain niveau de consensus, par exemple sur le langage ou sur l’ontologie utilisée. Il devrait contribuer à assurer autant d’automatisation que possible et à basculer entre différentes formes et ontologies. Il facilitera la mise en œuvre de calculs d'inférence complexes tout en offrant des garanties plus élevées quant à leur validité. Il doit fournir un mécanisme de protection, ainsi qu’un mécanisme d’éligibilité des connaissances pour augmenter le niveau de confiance de l’utilisateur .Dans le Web 1.0 ou Web documentaire on con state de nombreuses limites dans la recherche d'information ce qui conduit fréquemment au bruit ou silence documentaire. Un des objectifs de la mise en place du Web sémantique - ensuite appelé plus justement « Web des données » - est l’amélioration de la recherchée il est notamment possible de rechercher par concepts ou de trouver des informations d’une granularité plus fine. Dans ce contexte les Systèmes d’Organisation des Connaissances (SOC) occupent une place prépondérante puisqu’ils permettent de définir des termes ou concepts (et éventuellement des relatons entre eux) afin qu’un vocabulaire commun soit partagée par des usagers. [11]

.

2. L'évolution du web de documents au Web sémantique

2. 1. Le web documentaire

Le Web 1.0 a été créé par Tim Bernerns-Lee et Roboert Cailliau au CERN (Centre européen de recherche nucléaire). Ainsi à la fin de l'année 1990 lors du projet WebCore le premier serveur et le premier navigateur sont testés via une connexion internet. Web Documentaire, c'est-à-dire un ensemble de documents interconnectés accessibles sur Internet, destinés à permettre aux utilisateurs de naviguer d'un document à l'autre pour traiter eux-mêmes le contenu de l'information. Actuellement, le Web contient non seulement des documents, pour les agents humains, mais aussi des données pour les agents. Certaines données décrivent un document sur le Web tout en étant indépendantes de tout document sur le Web.

Le W3C a notamment formalisé les trois notions fondamentales à l'origine et qui constituent le cœur du Web documentaire :

- L'Universal Resource Identifier (**URI**) : c'est un format d'identifiants uniques permettant de nommer n'importe quelle ressource sur le Web. De plus si l'identifiant offre un chemin d'accès vers une représentation de la ressource c'est une Uniform Resource Locator (**URL**) ou adresse Web.
- L'HyperText Transfer Protocol (**HTTP**) : permet via une adresse URL d'accéder à une page Web identifiée et localisée par cette URL.
- HyperText Markup Language (**HTML**) : est un langage de balisage que l'on utilise pour représenter mettre en forme et publier des pages Web.

2.2. Vers le Web sémantique

Le terme « Web sémantique » est attribué à Berners-Lee et Lassila, qui le décrivent non pas comme différent de ce que nous connaissons, mais plutôt comme une extension de celui-ci. Il s'agit d'un ensemble de technologies dont le but est d'être plus efficace que la quantité d'informations importantes trouvées sur le Web. De grandes quantités de données structurées sont présentement conservées dans des bases de données isolées du Web, que l'on pourrait comparer à des silos d'informations, dont font partie les métadonnées des catalogues de bibliothèques. L'objectif est de les rendre accessibles et

.

de les encoder à l’aide de standards et de normes bien définis, de manière à ce que les machines puissent les interpréter et à assurer une meilleure collaboration entre humains et machines. Ainsi, la compréhension et l’interprétation du sens que prennent les mots et les métadonnées ne se limitent plus à l’humain. Cette interprétation permet une suite de gestion de l'information plus efficace, permettant la création de services et d’applications, et garantissant l'accès aux connaissances à la fois par les humains et les machines.

Les **4** principaux standards du Web sémantique :

- RDF: un modèle de triplets pour décrire et connecter des ressources anonymes ou identifiées par un URI
- SPARQL: un langage de requête sur les graphes RDF
- RDFS est un langage de descriptions légères
- OWL: 3 couches d'extension de l'expressivité (logique). [12]

3. Schéma de l'évolution du web de documents au Web des données liées (linked data) :

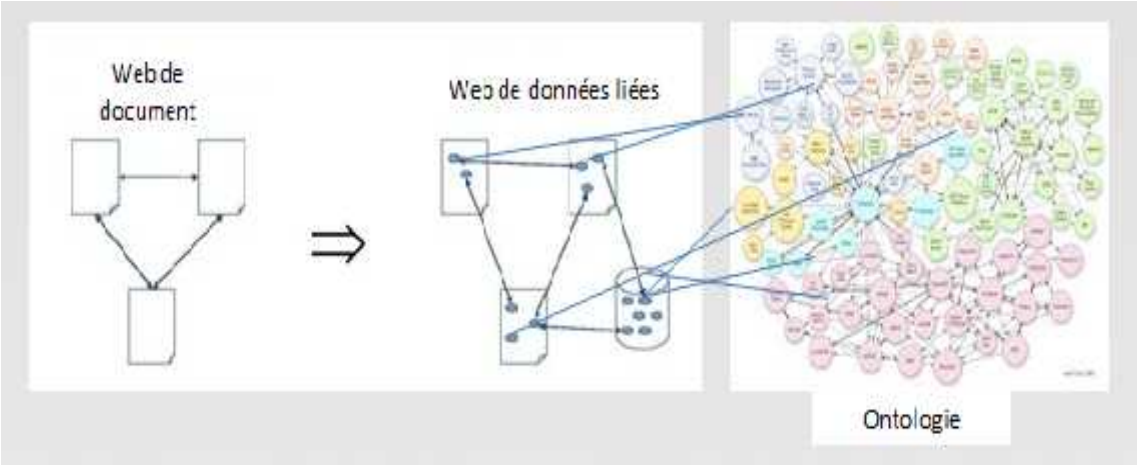


Figure 2 : L'évolution du Web traditionnel vers le Web sémantique.

L'évolution du Web traditionnel vers le Web sémantique permet à la machine d'être un citoyen de premier ordre sur le Web et augmente la découvrabilité et l'accessibilité des données non structurées sur le Web. [13]

.

4. Web Actuel Vs Web Sémantique

Web actuel	Web Sémantique (WS)
▪ Ensemble de **documents**	▪ Ensemble **d'information / connaissances**
▪ Basé essentiellement sur **HTML**	▪ Basé essentiellement sur **XML et RDF(S), OWL**
▪ Recherche par **mots-clés**	▪ Recherche par **concepts (ontologie)**
▪ Utilisable par **l'être humain**	▪ Utilisable par la **machine**

Tableau 5: Les principaux standards du Web sémantique et web de document.

5. Les architectures du Web Sémantique et web de document

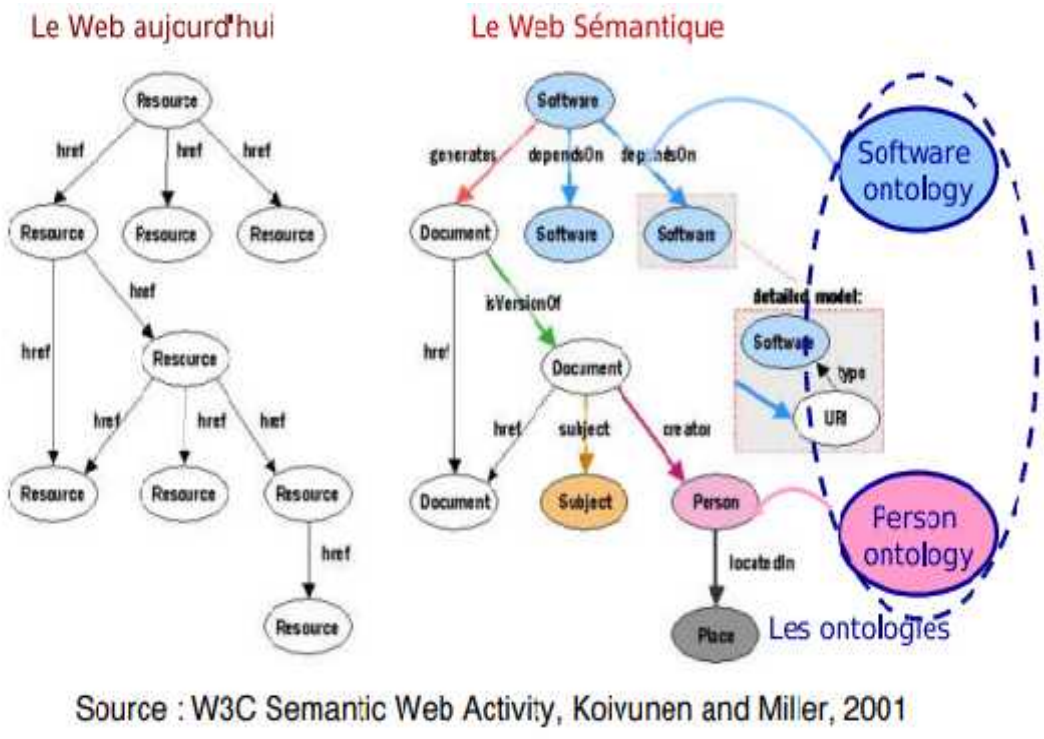


Figure3 : L'architecture du Web Sémantique et web de document.

6. Briques représentation : URL, URI, IRI

Objectif général : nommer les ressources

.

- URI : Uniform Resource Identifier (août 98) / IRI (International Resource Identifier) permet d'identifier une ressource (physique ou abstraite) sur le Web
- URL : URI qui donnent le moyen d'accéder à la ressource. Ex : http://www.wikipedia.org/
- URN : URI qui permettent d'identifier une ressource par son nom dans un espace de noms. Ex : urn:isbn:0-395-36341-1 [13]

8. Briques représentation : XML

Objectif général : représenter les ressources pour les machines

- XML : eXtensible Markup Language, recommandation XML 1.0 en février 98).
- XML : séparation fond/forme.
- Méta-langage, qui permet de définir des langages de documents.
- Nombreux dialectes : MathML, XSLT, XACML, SVG, XHTML, …
- Document valide : conforme à un schéma, défini par une DTD, un XMLschema, un schéma RelaxNG.
- Outils de manipulations, basés sur Xpath : XSLT, XQuery [13]

9. Briques représentation : RDF

Objectif général : exprimer les ressources et les relier

- RDF = la base du Web sémantique
- RDF = cadre pour décrire des données sur les ressources du Web (Resource Description Framework)
- Représentation des ressources par Triplets (Sujet, Prédicat (propriété), Valeur) :
    - Sujet : une ressource qui peut être identifiée par un URI
    - Prédicat : une spécification réutilisée et identifiée par URI de la propriété
    - Objet : une ressource ou constante à laquelle le Sujet est lié
- Triplets sérialisés de différentes façons (XML, Turtle, …)
- Permet de constituer des graphes RDF, des bases de données RDF (Triples-stores) [13]

.

10. Briques requêtes : SPARQL

Objectif général : faire des requêtes sur ressources exprimées en RDF (et RDF-S).

- SPARQL = Simple Protocol and RDF Query Language :
  - Un langage de requête pour RDF
  - Un protocole : spécification pour émettre et envoyer des requêtes SPARQL (services Web) vers des serveurs dédiés et en recevoir les résultats
  - Un format XML pour l'affichage des résultats obtenus (requêtes de type SELECT et ASK)
- Permet de réaliser des requêtes fines et précises
- Permet aussi de réaliser des opérations : ajout, modification, suppression, tris, ... de données RDF
- Inspiréde SQL pour la syntaxe et les fonctionnalités. [13]

.

11. Conclusion

Ce chapitre présente L'évolution du Web traditionnel vers le Web sémantique qui permet à la machine d'être un citoyen de premier ordre sur le Web et augmente la découvrabilité et l'accessibilité des données non structurées sur le Web. Cette évolution permet à la technologie Linked Data d'être utilisée comme base de connaissances de base pour les données non structurées, notamment les textes, disponibles aujourd'hui sur le Web. [4]

.

Chapitre 3 : Les ontologies.

1. Introduction

La masse de plus en plus croissante d'information dans tous les domaines a généré un besoin capital d'organisation et de structuration des contenus de documents, disponibles généralement sur le web. Les ontologies en sont un moyen prometteur et qui ne cesse de donner ses preuves. Leurs applications sont multiples : indexation, recherche d'informations, traduction automatique, eLearning etc.

Les principaux buts de la construction des ontologies sont la partageabilité, la portabilité, la réutilisabilité et la capitalisation de la connaissance et de l'expertise d'un domaine. Parce que l'information n'est pas statique, parce qu'elle se modifie, s'enrichisse, s'altère avec le temps et qu'elle vienne de différentes sources, nous avons besoin d'outils et de modèles qui permettent aux utilisateurs et aux experts du domaine de constituer, consulter et maintenir à jour leurs connaissances du domaine. [1]

2. Définitions

Le mot ***ontologie*** qui vient du grec ***ontos*** =être et ***logos***= études, appartient à la philosophie ancienne grecque, Aristote le définit comme la science de l'Être en tant qu'être [14]. Il est difficile de définir ce qu'est une ontologie d'une façon définitive. Le mot est en effet employé dans des contextes très différents touchant à la philosophie, la linguistique ou l'intelligence artificielle.

Bien que des débats préexistent, nous parlons plus souvent d'ontologies (au pluriel) afin de refléter les multiples facettes que recouvre cette appellation [15] abordent les différentes définitions de la littérature afin d'examiner le type de représentation des connaissances dénoté par le terme ontologie. En 1993, Gruber propose une première définition « *une ontologie est une spécification explicite d'une conceptualisation* », [16]. L'expression *spécification explicite* signifie, que la conceptualisation est représentée dans un langage qu'il soit naturel (Arabe, français...) ou formel (logique de description, graphes conceptuels..).

Une autre définition, peut-être plus rigoureuse : « *Une ontologie implique une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts (entités, attributs, processus, leurs définitions et leurs*

.

interrelations). On appelle cela une conceptualisation » [17]. Une ontologie peut prendre différentes formes mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification [18]. En résumé, nous pouvons définir une ontologie comme l'ensemble représentatif et exhaustif des termes d'un domaine donné avec toutes les relations qui les relient.

3. Types d'ontologies

Il existe de nombreuses classifications des ontologies selon des critères variés :

3.1. La classification faite par M. Uschold et M. Grüninger (1996)

Distingue les ontologies en fonction du degré de formalisme de la représentation :

- les ontologies hautement informelles sont des ontologies opérationnelles écrites en langage naturel ;
- les ontologies semi-informelles utilisent un langage naturel structuré et limité ;
- les ontologies semi-formelles définissent les concepts dans un langage artificiel et défini formellement ;
- les ontologies rigoureusement formelles sont définies dans un langage contenant une sémantique formelle, des théorèmes et des preuves de propriétés telles que la robustesse et l'exhaustivité.

3.2. La classification faite par A. Gómez-Pérez et al. (2004a) et N. Guarino (1998)

La classification peut également se faire en fonction des objets que modélisent les ontologies (comme illustré dans la Figure 4)

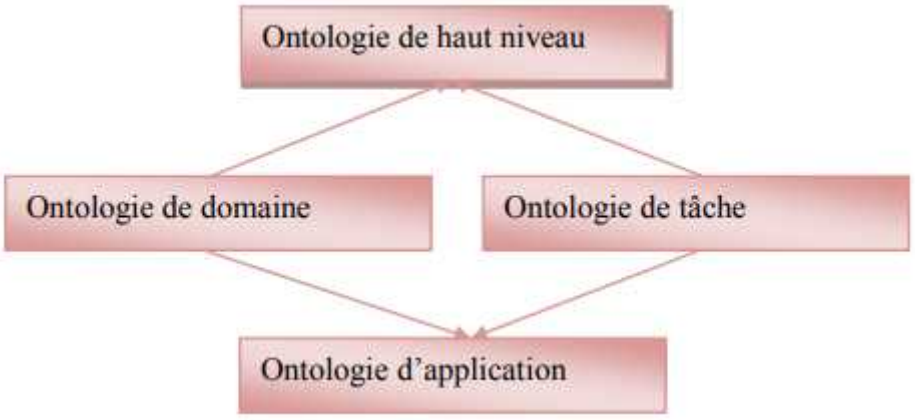


Figure 4 : Classification des ontologies selon N. Guarino (1998).

.

- L'ontologie de haut niveau ou ontologie de niveau supérieur est liée à des concepts de haut niveau qui décrivent des concepts très généraux. Ils conviennent aux grandes communautés d'utilisateurs et d'applications et peuvent être divisés en deux catégories :

- L'ontologie utilisée pour représenter la connaissance décrit les concepts utilisés pour spécifier la connaissance dans toute ontologie. Exemple : L'ontologie framework du projet Ontolingua3 [19] définit formellement les concepts principalement utilisés dans les langages framework : classes, sous-classes, attributs, valeurs, relations et axiomes ….
- L'ontologie générale (ou ontologie générale) décrit des concepts très généraux qui sont indépendants de domaines ou de problèmes spécifiques (espace, temps, événements). Ils peuvent être utilisés pour initier la construction d'une ontologie ou compléter une ontologie existante mais incomplète. Exemple : SUO Ontology [20] a été développé par SUO WG (Ontology Working Group on Standards).

- L'ontologie de domaine décrit la connaissance d'un domaine spécifique (voitures, hôpitaux, etc.). Les concepts et les relations de l'ontologie de domaine se réfèrent à des objets de domaine. Ils peuvent être obtenus en spécialisant les concepts de l'ontologie générale. La relation de ce type d'ontologie décrit les liens entre les concepts qui existent et sont valables dans le domaine considéré. Comme exemples d'ontologies de domaine, nous pouvons citer Menelas [21] ou Galen [22] dans le domaine médical, PhysSys [23] pour la Physique, ou encore les ontologies TOVE [24] [25] et Enterprise [26] dans le domaine de la Mémoire d'entreprise (domaine ayant fait l'objet de plusieurs recherches en Ingénierie des Connaissances [27] qui ont abouti au développement de méthodes et d'outils spécifiques de construction, de gestion et d'exploitation d'ontologies, comme par exemple la méthode SAMOVAR [28] pour la mémoire d'un projet de conception de véhicule) ;

- L'ontologie de tâche décrit les connaissances liées à une tâche ou une activité spécifique (vente, navigation, etc.). Comme l'ontologie de domaine, le concept d'ontologie de tâche peut être obtenu en spécialisant le concept d'ontologie générale. L'ontologie de tâche clarifie le rôle de chaque concept dans l'activité de modélisation. Plusieurs ontologies de tâches ont été développées dans le cadre du projet Ontolingua.

.

- L'ontologie appliquée est l'ontologie la plus spécifique. Ils donnent des concepts spécifique pour un et pour une activité particulière. L'ontologie appliquée peut être considérée comme un double de l'ontologie du domaine et de l'ontologie des taches. Ici, les concepts qui décrivent généralement les objets de domaine avec une opération spécifique. On peut citer les travaux sur la modélisation de ressources pour des applications d'e-formation [29] [30]. Les ontologies d'application sont en général utilisées pour élaborer des applications concrètes mais ne doivent pas être confondues avec des bases de connaissances.

4. Rôle des ontologies dans le Web Sémantique

- Définir de manière déclarative un vocabulaire commun résultat d'un consensus social dans un domaine donné :
  - Chaque élément de vocabulaire possède une interprétation unique partagée par tous les membres du domaine
- Décrire la sémantique des termes et leurs relations :
  - L'interprétation de chaque terme est unique er résulte d'une sémantique formelle.
  - L'ensemble des termes et leurs relations fournissent un cadre interprétatif dépourvu d'ambiguïté pour chaque terme.
- Fournir des mécanismes d'inférence qui respectent la sémantique formelle. [31]

5. Langage d'ontologies

Langages de description d'ontologie du W3C :

5.1. RDF-Schéma :

- Permet de décrire un vocabulaire RDF spécifique à un domaine (RDF vocabulary description language)
- Fournit une sémantique à ce vocabulaire en décrivant les propriétés et les classes des ressources RDF

.

- Utilisé pour formaliser des ontologies légères en permettant un raisonnement limité sur ces ontologies [13]

5.2. OWL:

Web Ontology Language (issu de DAML+OIL)

- Langage plus puissant de formalisation d'ontologies : relations entre classes, contraintes de cardinalité propriété de typage plus riches, …
- Utilisé pour formaliser des ontologies lourdes en permettant un raisonnement puissant sur ces ontologies en s'appuyant sur les logiques de description (LD). [13]

.

6. Conclusion

Nous avons abordé dans ce chapitre un tour d'horizon sur les différentes technologies utilisées dans le développement d'ontologies. Nous n'avons présenté qu'une liste succincte mais nous avons tenu à ce qu'elle soit la plus représentative possible des outils existant pour chaque phase de la création d'ontologie. Nous avons mis l'accent autant que cela était possible sur les outils libres et open source, qui peuvent servir beaucoup plus, dans le domaine de la recherche. [1]

Chapitre 4: Annotation Automatique des textes arabes avec Linked Data.

1. Introduction

Dans cet article, j'introduis une approche à base des règles pour l'annotation automatique des expressions de la localisation et de la direction en arabe. La visée de cet article est double : il s'agit premièrement de définir les structures linguistiques qui expriment la localisation et la direction en arabe, et deuxièmement d'envisager le problème de l'annotation automatique des segments textuels conformes à ces structures linguistiques.

Je présente un système d'annotation des connaissances spatiales en arabe, en se basant sur une analyse linguistique profonde et sur une carte sémantique du domaine de la spatialité.

L'objectif est proposé une approche pour l'annotation de textes arabes avec Linked Data, notamment la base de connaissances DBpedia. Cela constitue la première étape pour permettre aux utilisateurs de lier des documents texte au cloud DBpedia Linked Open Data via l'environnement Web sémantique. [32]

2. L'annotation :

Diverses définitions ont été données pour le terme annotation. Ces définitions sont spécifiques à une opinion et ne sont pas tout à fait cohérentes avec notre recherche. "Une annotation est une information graphique ou textuelle attachée à un document et souvent placée dans ce document" [33]. Cette définition est certainement valide, mais la relation entre source et annotation n'est pas claire. Cependant, pour nous, l'association entre le document source et l'annotation est obligatoire. S'il n'y a pas de lien ou de lien entre eux, on ne peut pas parler d'annotation. L'association doit être déclarée directement ou indirectement. "Une annotation est une explication qui accompagne un texte" dit le Petit Robert, dictionnaire de la langue française Dans cette définition le mot d'accompagnement est important, plutôt formel, car il est clair que toute annotation, n'existe qu'en relation avec "texte" - En revanche, comme on peut

reconnaître que le mot n'a qu'un seul sens, cette définition exclut les annotations appliquées aux objets graphiques ou aux systèmes sonores.

L'annotation de ressources documentaires est une vieille tradition dans le monde de la documentation et des bibliothèques. La Digital Library Federation4 (DLF), une association constituée des quinze bibliothèques américaines les plus importantes aux Etats-Unis, a défini trois sortes d'annotations qui peuvent s'appliquer aux ressources documentaires d'une bibliothèque numérique :

- l'annotation administrative, ou, qui indique des informations sur la création et la maintenance du documentaire, telles que « qui, quoi, où et comment ». Depuis l'avènement du Web, le langage DublinCore a servi de standard pour les annotations avec des descripteurs tels que l'auteur, le titre, la source, l'éditeur, la date de publication, etc.
- L'annotation structurelle relie des parties d'un parties d'un document source ensemble pour former une présentation logique du document.
- L'annotation descriptive décrit le contenu visuel du document source, c'est-à-dire qu'elle identifie les concepts couverts dans la source, les relations entre ces concepts et les instances. [33]

3. Travaux connexe

L'arabe exprime des informations locatives et positionnelles très riches et intéressantes. Le travail se situe à la croisée de deux grandes perspectives :
(1) une perspective linguistique, d'une part, et (2) la perspective informatique, d'autre part.
Sur le plan linguistique, l'objectif initial de ce travail est de présenter un essai de classification générale des prépositions de localisation et de la direction en prenant en compte leurs propriétés sémantiques.
Le deuxième objectif, sur le plan informatique, concernait l'écriture de règles informatiques, implémentables dans la plateforme NooJ, permettant d'extraire automatiquement une expression de localisation dans un corpus arabe.

De nombreux travaux se réclament des grammaires cognitives de [34] et [35], [36] [37], [38] [39], [40] et [41] (pour ne citer que quelques références), qui portent sur la cognition de l'espace et sur la sémantique des prépositions spatiales.

Notre approche se base sur des études faites sur d'autres langues. Dans son étude, Kopecka, présente la typologie de l'expression de la localisation et du déplacement en français et en polonais dont le but était d'explorer l'expression de ces domaines sémantiques dans les deux langues dans une perspective typologique afin d'évaluer son impact sur l'élaboration linguistique de l'information spatiale. [42]

En arabe, Mubarak, dans son livre fournit une analyse détaillée des effets de sens de plusieurs prépositions, sans se limiter à leurs emplois spatiaux. Il s'intéresse à étudier la sémantique et l'usage de chaque préposition. [43]

L'intérêt particulier du choix de la langue arabe repose sur la rareté des travaux élaborés dans cette langue qui traitent ce sujet. En addition, dans notre étude, nous voulons montrer que l'ingénierie des langues peut exécuter les phénomènes de localisation et de direction spatiale. De plus, montrer que l'interaction avec les données décrivant des entités spatiales est possible afin de les enrichir d'une façon semi-automatique tout en se basant sur des outils informatiques déjà développés. Ce travail est l'un de nos travaux surl'annotation sémantique de textes en langue arabes (Alhajj et [44] [45].

Dans cet article nous essayons de présenter un système d'annotation pour l'arabe qui, après avoir effectué une approche linguistique pour annoter le corpus textuel arabe avec Linked Data, en particulier DBpedia, qui est Linked Open Data (LOD) extrait de Wikipédia. Cette approche utilise des techniques de langage naturel pour éclairer le texte arabe avec Linked Open Data.

4. Description du système

Notre système prend en entrée un texte arabe et fournit un texte étiqueté avec les URI de données DBpedia et Wiki, et il se compose de deux modules: (1) module de ressources du candidat et (2) module sémantique. Le premier permet de télécharger du texte arabe, de le diviser en phrases et de procéder au traitement de chaque phrase à l'aide du module d'extraction de ressources pour extraire toutes les ressources candidates. Le deuxième module vérifie l'existence de chaque ressource dans la base de connaissances. Enfin, le système produit un texte étiqueté à l'aide de la technologie Linked Data.

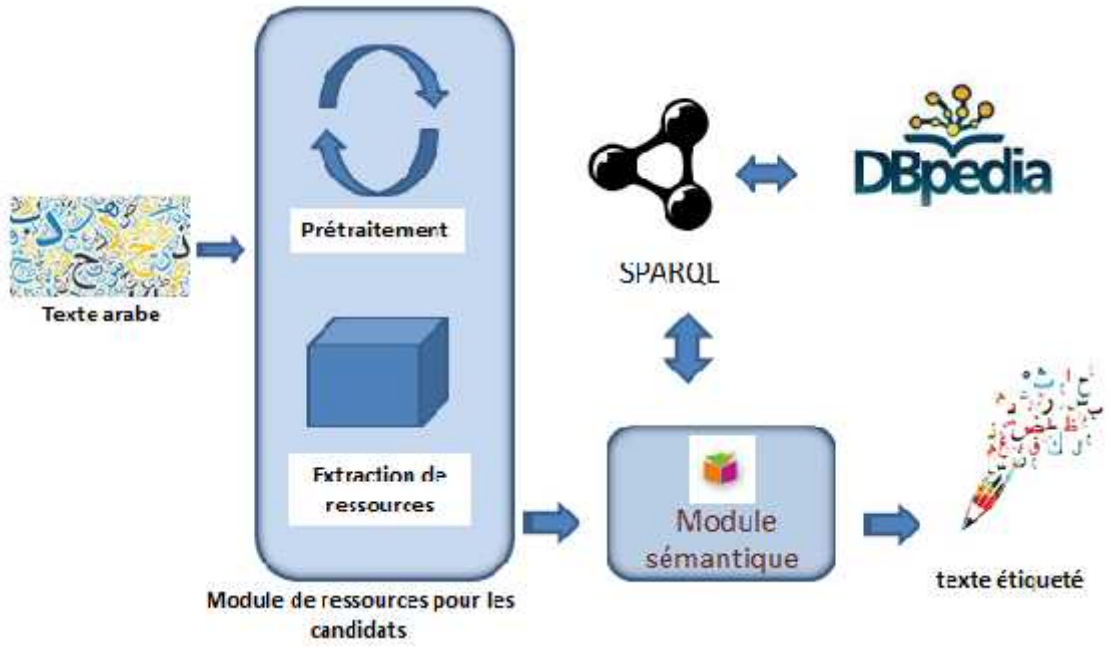


Figure 4 :l'architecture du système proposé.

J'ai détaillé chaque module et étape de notre architecture système proposée comme de suit :

4.1. Module de ressources pour les candidats

Les phrases peuvent être considérées comme des paires de ressources liées entre elles à l'aide de verbes, de prépositions, de conjonctions, de phrases verbales ou de phrases nominales. Dans les phrases, l'ensemble des mots considérés comme ressources

candidates sont ceux que l'on retrouve dans l'ontologie. Le processus de génération de ressources candidates est le suivant:

4.1.1. Prétraitement

Les étapes courantes de la plupart des processus NLP sont la tokenisation et la normalisation. La tokenisation désigne la segmentation d'un texte en langage naturel en unités de base consécutives individuelles. Une étape de normalisation des mots est nécessaire lorsqu'il s'agit de la langue arabe. La correction des fautes d'orthographe les plus courantes implique la normalisation des lettres: 'أ />', 'إ / <' et 'آ / |' sont remplacés par «ا / A», tandis que la lettre «ة/ p» est remplacée par «ه/ h», et la lettre «ى / Y» est remplacée par «ي/ y» [46]. Les signes diacritiques sont également supprimés à ce stade. Ces erreurs apparaissent dans le texte lorsque les rédacteurs ne respectent pas les règles grammaticales standard de l'arabe, et par conséquent, certaines lettres sont écrites dans des styles différents.

Notez que tout au long de cet article, chaque fois que nous donnons un texte arabe et pour des raisons de lisibilité, nous le suivons avec sa translittération buckwalter [47] et, éventuellement, traduction anglaise à des fins de lisibilité.

4.1.2. Extraction de ressources

Ce module vise à extraire toutes les ressources candidates du texte d'entrée. Pour cela, nous supposons que les ressources, existant dans les phrases, se présentent sous la forme de phrases nominales, ou d'une séquence de différentes formes de noms, noms propres et adjectifs, comme on peut le voir dans le tableau suivant.

Nominal phrase	Example
Propoer noun/st of proper nouns	**مصطفى، محمد علي/ mSTfY, mHmd Ely\muSTafaY, muHamed Ealiy**
Noun/set of nouns	**قصر المرادية/ qSr AlmrAdyp \qaSru AlmurAdiyap**
Noun +adjective	**المدنية الجديدة/ Almdynp Aljdydp \Almadynapu Aljdydap/**
Set of nominal phrases	**المدينة الجديدة سيدي عبد الله /Almdynp Aljdydp Sydy Ebd Allh \Almadynapu Aljadydap sydy Eabda Allah**

Tableau 5: Différentes formes grammaticales des ressources.

Le triple RDF <sujet, prédicat, objet> est le fait de représentation standard dans la technologie Linked Data. Le sujet et l'objet du triplet RDF sont généralement nommés avec des phrases nominales et peut-être des classes, des instances ou des valeurs littérales. [48]

4.2. Module sémantique

Après avoir extrait la liste des ressources candidates comme des informations précieuses à partir de phrases arabes, nous devons trouver la référence de chaque ressource qui peut correspondre à une ressource DBpedia. Cependant, l'ontologie DBpedia est en anglais et le chapitre arabe de DBpedia [49] n'est plus disponible. Pour ces raisons, nous utilisons le RDFS: label qui est une instance de RDF: property. rdfs: label est utilisé pour fournir une version lisible par l'homme dans une langue différente du nom d'une ressource. Dans la ressource, le Web of Data a de nombreux URI équivalents. owl: sameAs fournit un service pour trouver une coréférence entre différents ensembles de données dans les données liées.

Pour étiqueter et étendre les ressources dans les phrases, nous utilisons une requête SPARQL pour vérifier l'existence des ressources candidates dans DBpedia, puis interconnectons chacune avec cette base de connaissances. Nous utilisons la requête SPARQL suivante via le service en ligne de DBpedia SPARQL:

```
PREFIX owl: http://www.w3.org/2002/07/owl#
 PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
   PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
   PREFIX foaf: <http://xmlns.com/foaf/0.1/>
   PREFIX dc: <http://purl.org/dc/elements/1.1/>
   PREFIX : <http://dbpedia.org/resource/>
   PREFIX dbpedia2: <http://dbpedia.org/property/>
   PREFIX dbpedia: <http://dbpedia.org/>
   PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
   PREFIX dbo: <http://dbpedia.org/ontology/>
   SELECT ?s ?o
   WHERE { ?s rdfs:label resource.
               ?s owl:SameAs ?o.          }
```

Figure5 : Requête SPARQL renvoyant l'URI d'une ressource arabe en utilisant rdfs: label

La réponse du point de terminaison SPARQL est exprimée dans un fichier JSON. Après avoir analysé ce fichier JSON, nous extrayons les ressources texte et les URI d'entités similaires dans la base de connaissances. Ensuite, le système procède à lier les mots de la phrase (ressources) avec les URI pertinents dans la base de connaissances DBpedia.

5. Exemple illustratif

Dans cette section, nous donnons un exemple montrant comment notre système annote un texte arabe avec des données liées:

5.1. Texte arabe

هنأت موسكو وواشنطن وعواصم عربية أخرى أمس الجمعة الجزائر بإجراء الانتخابات الرئاسية التي فاز بها المرشح عبد المجيد تبون.'/hn>t mwskw wwA$nTn wEwASm Erbyp >xrY >ms AljmEp AljzA}r b<jrA' AlAntxAbAt Alr}Asyp Alty fAz bhA Almr$H Ebd Almjyd tbwn

\hana>at mwsokw wwaA$inTun waEawASim Earaboyap >uxraY >ams AljumuEap AljazA}ir bi<ijrA' AlAintixAbAt Alri}Asiyap Al~ty fAza bihA Almura$~H Eabdu Almajyd tabwn\

5.2. Résultat du Pre-processing

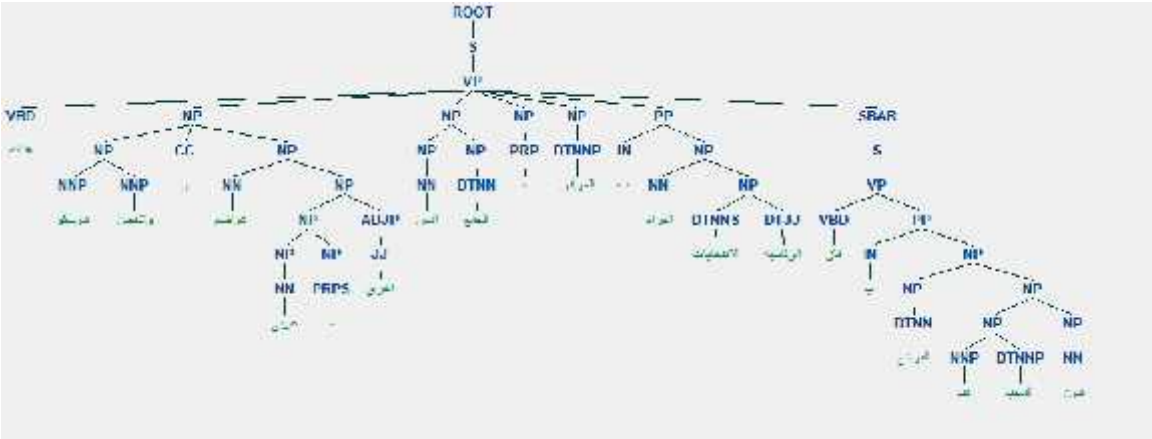
هنات موسكو و واشنطن و عواصم عربية اخرى امس الجمعة الجزائر ب اجراء الانتخابات الرئاسية التي فاز ب ها المرشح عبد المجيد تبون/ hnAt mwskw wA$nTn w EwASm Erbyp Axr Ams AljmEp AljzA}r b AjrA' AlAntxAbAt Alr}Asyp fAz b Almr$H Ebd Almjyd tbwn \hnAt mwskw w wA$nTn w EwASm Erbyp AxrY Ams AljmEp AljzA}r b AjrA' AlAntxAbAt Alr}Asyp Alty fAz b hA Almr$H Ebd Almjyd tbwn\

['هنات', 'موسكو', 'و', 'واشنطن', 'عواصم', 'عربية', 'اخرى', 'امس', 'الجمعة', 'الجزائر', 'ب', 'اجراء', 'الانتخابات', 'الرئاسية', 'التي', 'فاز', 'ب', 'ها', 'المرشح', 'عبد', 'المجيد', 'تبون']

5.3. Résultat du POS-tagging

[('هنات', 'VBD'), ('موسكو', 'NNP'), ('واشنطن', 'NNP'), ('و', 'CC'), ('عواصم', 'NN'), ('عربي', 'NN'), ('ه', 'PRP$'), ('اخري', 'JJ'), ('امس', 'NN'), ('الجمع', 'DTNN'), ('ه', 'PRP'), ('الجزائر', 'DTNNP'), ('ب', 'IN'), ('اجراء', 'NN'), ('الانتخابات', 'DTNNS'), ('الرئاسيه', 'DTJJ'), ('فاز', 'VBD'), ('ب', 'IN'), ('المرشح', 'DTNN'), ('عبد', 'NNP'), ('المجيد', 'DTNNP'), ('تبون', 'NN')]

5.4. Résultat du Parsing



5.5. Ressources pour les candidats:

["'موسكو ولشطن'@ar", "'عواصم عربية'@ar", "'الجمعة'@ar", "'اجراء الانتخابات الرئاسية'@ar", "'الانتخابات'@ar", "'الجزائر'@ar", "'ولشطن'@ar", "'موسكو'@ar", "'عبد المجيد'@ar", "'عبد المجيد تبون'@ar", "'الرئاسية'@ar", "'المجيد'@ar",]

5.6. Ressources en arabe annotées avec les URI DBpedia

[['http://dbpedia.org/resource/Friday', '"الجمعة"@ar'],

['http://www.wikidata.org/entity/Q262', '"الجزائر"@ar'],

['http://www.wikidata.org/entity/Q858439', '"الانتخابات الرئاسية"@ar'],

['http://www.wikidata.org/entity/Q500672', '"المرشح"@ar'],

['http://www.wikidata.org/entity/Q649', '"موسكو"@ar'],

['http://www.wikidata.org/entity/Q761885', '"ولشنطن"@ar'],

6. Implémentation

Nous avant utilise python pour l'implémentation de notre projet pour cela on a utilisé les bibliothèques suivantes :

6.1. Prétraitement

```
def clean_str(text):
    search = ["Ø£","Ø¥","Ø¢","_","-","/",".","ØŒ"," Ùˆ "," ÙŠØ§ ","”","Ù€","”","Ù‰","\\","\n', '\t','&quot;','؟','ØŸ','!']
    replace = ["Ø§","Ø§","Ø§","Ù‡"," "," ","","","","","," Ùˆ"," ÙŠØ§","","","","ÙŠ","","' ',' ',' ',' ؟ ',' ØŸ ',' ! ']

    #remove tashkeel
    p_tashkeel = re.compile(r'[\u0617-\u061A\u064B-\u0652]')
    text = re.sub(p_tashkeel,"", text)

    #remove longation
    p_longation = re.compile(r'(.)\1+')
    subst = r"\1\1"
    text = re.sub(p_longation, subst, text)

    text = text.replace('Ùˆ', 'Ùˆ')
    text = text.replace('ÙŠÙŠ', 'ÙŠ')
    text = text.replace('Ø§Ø§', 'Ø§')

    for i in range(0, len(search)):
        text = text.replace(search[i], replace[i])

    #trim
    text = text.strip()

    return text
```

6.2. Extraction des ressources

```
def extract_nps(tree):
 leav=list()
 nps=list()
 for t in tree.subtrees():
    if (t.label()=="NP"):
     leav.append(t.leaves())
 for tt in leav:
  te=""
  for i in range(0,len(tt)):
   te=te+tt[i]+" "
  nps.append(te[0:len(te)-1])
 nps=[w for w in nps if len(w)>2]
 n=[ns for ns in nps if (len(ns.split(" "))>1) and (len(ns.split(" "))<4)]
 nps = list(dict.fromkeys(n))
 return nps
```

6.3. Normaliser la ressource

```
def normalise(nps):
 res=list()
 a=""
 for n in nps:
  a="'"+n+"'@ar"
  res.append(a)
 return res
```

6.4. Module sémantique

```
def sparql_end(r):
    query="""
    PREFIX owl: <http://www.w3.org/2002/07/owl#>
    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
    PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX foaf: <http://xmlns.com/foaf/0.1/>
    PREFIX dc: <http://purl.org/dc/elements/1.1/>
    PREFIX : <http://dbpedia.org/resource/>
    PREFIX dbpedia2: <http://dbpedia.org/property/>
    PREFIX dbpedia: <http://dbpedia.org/>
    PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
    PREFIX dbo: <http://dbpedia.org/ontology/>
    SELECT distinct ?s ?type
    WHERE { ?s rdfs:label """+r+""". }"""
    #?type rdfs:subClassOf owl:Thing ?s rdf:type ?type .
    sparql = SPARQLWrapper("http://dbpedia.org/sparql")
    sparql.setQuery(query)
    sparql.setReturnFormat(JSON)
    results = sparql.query().convert()
    #for result in results["results"]["bindings"]:
     #print('%s: %s' % (result["label"]["xml:lang"], result["label"]["value"]))
    return results
```

6.5. Extraction des entités dbpedia

```
def uri_db(res):
 uri=list()
 for r in res:
    l=list()
    s=sparql_end(r)
    if (s['results']['bindings']!=[]):
     l.append(s['results']['bindings'][0]['s']['value'])
     l.append(r)
     uri.append(l)

 return(uri)
```

7. Expériences et évaluation

Pour évaluer notre système proposé, nous avons mis en place un outil prototype et nous avons réalisé des expériences à l'aide d'un corpus textuel de 100 phrases extraites d'un journal en ligne couvrant différents domaines ; ce corpus a été annoté manuellement à l'aide des URI DBpedia. Notre choix est argumenté par l'absence d'une référence dorée pour les corpus de textes annotés en arabe.

Notre système a été évalué à l'aide des mesures de précision, de rappel et de mesure F définies comme suit :

Metric	Definition
Precision	$\frac{\text{Correctly annotated information}}{\text{Total number of the annotated information (by our system)}}$
Recall	$\frac{\text{Correctly annotated information}}{\text{Total number of genrated annotation information (by our sytem)}}$
F-measure	$2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$

Tableau 7 : Paramètres d'évaluation des processus d'extraction de ressources et d'interconnexion

8. Les résultats globaux de l'évaluation

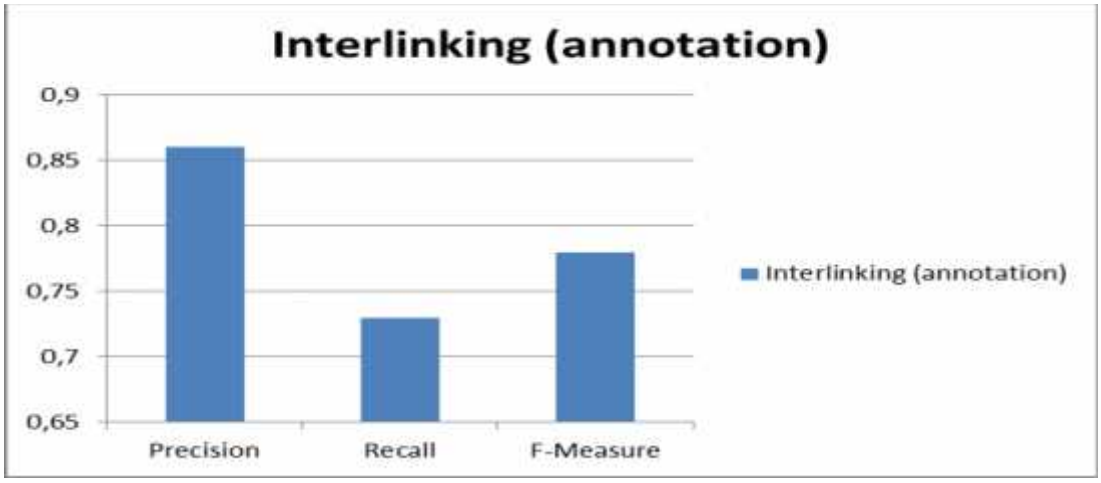


Figure 7: Résultats de l'évaluation du système d'annotation.

Les résultats sont très encourageants et prometteurs car l'annotation atteint une précision de 0,86, le rappel de 0.73 et la mesure F de 0,79.

9. Conclusion:

Ce chapitre présente la description du notre système qui prend en entrée un texte arabe et fournit un texte étiqueté avec les URI de données DBpedia et Wiki. Notre système se compose de deux modules: (1) module de ressources du candidat et (2) module sémantique. Le premier permet de télécharger du texte arabe, de le diviser en phrases et de procéder au traitement de chaque phrase à l'aide du module d'extraction de ressources pour extraire toutes les ressources candidates. Le deuxième module vérifie l'existence de chaque ressource dans la base de connaissances. Enfin, le système produit un texte étiqueté à l'aide de la technologie Linked Data.

Conclusion générale

Dans ce mémoire, nous avons proposé une approche pour annoter des textes arabes avec des données liées, notamment la base de connaissances DBpedia. Cela constitue la première étape pour permettre aux utilisateurs de lier des documents texte au cloud DBpedia Linked Open Data via l'environnement Web sémantique.

Le Web sémantique, en tant qu'extension du Web traditionnel, est moins riche en ressources arabes. Ainsi, plus d'efforts doivent être faits dans ce domaine pour combler cette lacune et permettre à la langue arabe de satisfaire les besoins des utilisateurs arabes sur le Web.

Les résultats de l'évaluation montrent l'efficacité du système proposé, même dans un environnement Open Linked Data qui n'a pas de domaine contextuel de discours et a un vocabulaire hétérogone. Dans le contexte de bases de connaissances à domaine restreint, telles que médicales, militaires, juridiques et législatives, etc., dans lesquelles le vocabulaire est limité, les résultats doivent être plus intéressants. Les résultats peuvent également être améliorés avec davantage de techniques NLP basées sur des algorithmes d'apprentissage automatique, notamment la reconnaissance d'entités nommées. [4]

Références et bibliographie

[1]Soraya Zaidi–Ayad, « Une plateforme pour la construction d'ontologie en arabe : Extraction des termes et des relations à partir de textes (Application sur le Saint Coran) », Thèse, Faculté des Sciences de l'Ingénieur, Université BADJI MOKHTAR-ANNABA, 2012/2013.

[2]Abbes, R. (2004). *La conception et la réalisation d'un concordancier électronique pour l'arabe*. *Thèse de Doctorat*, L'institut national des sciences appliquées de Lyon.

[3]Noureddine Doumi, « Extraction d'information à partir d'un texte arabe: extraction des entités nommées et leurs relations sémantiques ». Intelligence artificielle [cs.AI]. Université Djillali Liabes de Sidi Bel Abbès, 2017. Français.

[4]Abdelghani BOUZIANE et all, «Annotating Arabic Texts with Linked Data».Ctr Univ Naama, Inst. Sciences and Technologies, Dept. Mathematics and Computer Science, EEDIS Lab.,UDL-SBA, Algeria.

[5]Hadj henni M. (2007). *Approche ontologique pour la modélisation sémantique, l'indexation et l'interrogation des documents Coraniques*, Mémoire de Magister, Ecole Supérieur d'Informatique (E.S.I) Alger.

[6]Jarrar M., Ayesh S., Al-Badawi M., Samara H. (2010). *Towards Building An Arabic Ontology*. Technical Report, Faculty of Information Technology, Birzeit University.

[7]KHODJA Ala Eddine, « Un système d'extraction d'information pour la langue arabe ». FACULTE: Mathématique et Informatique. UNIVERSITE MOHAMED BOUDIAF - M'SILA. 2016/2017.

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

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

[10]LAKEL Kheira, « Les annotations sémantiques dans les documents Web : application aux textes psychologiques en langue arabe ». *Faculté : Mathématiques et Informatique*. Universitaire Sciences and Technologies_Mohammed Boudiaf d'Oran .2017/2018

Table des abréviations.

[11] Berners-Lee Tim, Hendler James & Lasilla Ora (2001). The Semantic Web, Scientific American,.

[12]Marielle St-Germain. « Le Web de données et le Web sémantique à Bibliothèque et Archives nationales du Québec : constats et recommandations fondés sur l'initiative de la Bibliothèque nationale de France ». Mémoire présenté à la Faculté des études supérieures et postdoctorales en vue de l'obtention du grade de maître en Sciences de l'information. Marielle St-Germain, 2016

[13]Bernard ESPINASSE. « Introduction au Web Sémantique ». *Aix-Marseille Université LIS UMR CNRS 7020*. Septembre 2019.

[14]Welty C., & Guarino, N., (2001). *Supporting Ontological Analysis of Taxonomic Relationships* Data et Knowledge Engineering (39), pages 51-74, 2001.

[15]Baneyx A., (2007). *Construire une ontologie de la pneumologie, aspects théorique, modèles et expérimentations*. Thèse de doctorat, Université Pierre et Marie Curie.

[16]Gruber T. (1993). A *translation approch to portable ontology specifications*. Knowledge acquisition, 5(2), 199–220.

[17] Charlet J., Bachimont B., Bouaud J., Zweigenbaum P., (1996). *Ontologie et réutilisabilité :expérience et discussion*. In N. Aussenac-Gilles, P. Laublet & C. Reynaud, Coordinateurs, Acquisition et ingénierie des connaissances : tendances actuelles, chapitre 4, p. 69–87. Cepaduès-éditions.

[18]Baneyx A., & Charlet, J., (2006). *Évaluation, évolution et maintenance d'une ontologie en médecine: état des lieux et expérimentation*, Revue I3 ; SI 2006 special issue on Ontological ressources.

[19]GRUBER T. (1993). A translation approch to portable ontology specifications. *Knowledge Acquisition, 5(2), 199–220.*

[20] Niles, I., and Pease, A. (2001). Origins of the Standard Upper Merged Ontology: A Proposal for the IEEE Standard Upper Ontology. *In Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology, Seattle, Washington, August 6, 2001*.

[21] P. Zweigenbaum, B. Bachimont, J. Bouaud, J. Charlet et J.-F.Boisvieux. (1995). Issues in the structuration and acquisition of an ontology for medical language understanding, *Methods of Information in Medicine, Vol. 34(1/2), pp. 15-24, 1995.*

[22] A. L. Rector, W. D. Solomon, W. A. Nowlan, et T. W. Rush, A. (1994). Terminology Server for Medical Language and Medical Information Systems, *Methods of Information in Medicine, Vol. 34, pp. 147-157, 1994.*

[23] W. N. Borst, J. M. Akkermans, et J. L. Top, (1997). Engineering Ontologies, *International Journal of Human Computer Studies (Special Issue on Using Explicit Ontologies in KBS Development), Vol. 46, pp. 365-406, 1997.*

[24] M. S. Fox (1992). The TOVE Project: A Common-sense Model of the Enterprise, Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, *F. Belli et F. J. Radermacher (réd.), Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence subseries), LNAI 604, Springer, pp. 25-34, 1992.*

[25] H. M. Kim, M. S. Fox et M. Gruninger (1999). An ontology for quality management – enabling quality problem identification and tracing, *BT Technology Journal, Vol. 17, n° 4, pp. 131-140, 1999.*

[26] M. Uschold, M. King, S. Moralee et Y. Zorgios (1998), The Enterprise Ontology. The Knowledge Engineering Review, Special Issue on Putting Ontologies to Use, *M. Uschold et A. Tate (réd.), Vol. 13, 1998.*

[27]R. Dieng-Kuntz, O. Corby, F. Gandon, A. Giboin, J. Golebiowska, N. Matta et M. Ribière(2001), Méthodes et outils pour la gestion des connaissances : une approche pluridisciplinaire du Knowledge Management, *Collection Informatiques (Série Systèmes d'information), Dunod, France, ISBN 210006300-6, 2001.*

[28] J. Golebiowska (2002), Exploitation des ontologies pour la mémoire d'un projet-véhicule : Méthode et outil SAMOVAR, *Thèse de Doctorat (Spécialité Informatique), INRIA, Université de Nice-Sophia Antipolis, France, 2002.*

[29] A. Benayache, C. Barry, B. Chaput et M.-H. Abel (2004), Construction of application ontology for e-learning, *Proceedings of the World Conference on E-Learning in Corporate, Government Healthcare & Higher Education (E-Learn04), Washington,USA, 2004.*

[30] B. Chaput, A. Benayache, C. Barry et M.-H. Abel (2004), Une expérience de construction d'ontologie d'application pour indexer les ressources d'une formation en statistique, *Actes des Trente-sixièmes Journées Françaises de Statistique (SFDS'04), Montpellier, France, 2004.*

[31]MAACHE SALAH. « Apport des Ontologies pour les Placements Financiers Intelligents ». Vice rectorat chargé de la post graduation Ecole doctorale d'informatique. UNIVERSITE FERHAT ABBES SETIF
Année 2011.

[32]Rita Hijazi, *A.Sabra, M.Al-Hajj* « Annotation Automatique Des Connaissances Spatiales en arabe ». Université Libanaise, Centre des Sciences du Langage et de la Communication, Faculté des Lettres, Tayouneh, Beyrouth, Liban, 2018.

[33]Charles Robert. » L'annotation pour la recherche d'information dans le contexte d'intelligence économique ». domain_stic.docu. Université Nancy II, 2007. Français.

[34]Jackendoff R. (1991). « Parts and boundaries », Cognition 41, 9-45.

[35] Jackendoff R. (1996). « The Architecture of linguistic–spatial interface », in BLOOM P., PETERSON M.A., NADEL L. & GARRETT M.F. (eds), Language and Space, Cambridge (Mass.), The MIT Press, 2-32.

[36] Herskovits A. (1986). Language and Spatial Cognition. An Interdisciplinary Study of Prepositions in English. Cambridge : Cambridge University Press.

[37] Herskovits A. (1997). Language, spatial cognition, and vision. In O. Stock (ed.), Spatial and temporal reasoning, Dordrecht : Kluwer Academic Publishers, 155-202.

[38] Vandeloise C. (1986). L'espace en français : sémantique des prépositions spatiales, Paris : Seuil.

[39] Vandeloise C. (1992). Les analyses de la préposition dans : Faits linguistiques et effets de méthodologie, Lexique 11, 15-40.

[40] Vandeloise C. (1999). Quand dans quitte l'espace pour le temps. Approches sémantiques des prépositions, Revue de Sémantique et Pragmatique 6, 145-163.

[41] Talmy L. (2000). Toward a Cognitive Semantics, Cambridge (Mass.), The MIT Press.

[42] Kopecka A. (2004). Etude typologique de l'expression de l'espace ; Localisation et déplacement en français et en polonais. Thèse de doctorat, Université Lumière, Lyon2.

[43] Mubarak, Abd Hussein (1988), huruf algr wamaḏahib alnuḥat fi istʿmaliha. Available on: http://qspace.qu.edu.qa/bitstream/handle/10576/8915/028811-0005fulltext.pdf?sequence=10

[44] Alhajj M, Mourad G. (2015). Extraction of reported speeches from Arabic Lebanese newspapers. IEEE The Fifth International Conference on Digital Information and Communication Technology and its Applications (DICTAP2015), Faculty of Engineering - Lebanese University, Lebanon, April 29 - May 1, 2015 Proceedings.

[45] Alhajj M, Sabra A., (2018). Automatique Identification of Arabic Expressions Related to Future Events in Lenanon's Economy. International Journal of Science and Research (IJSR), volume 7, Issue 4.

[46]Exner, P. and P. Nugues (2012). Entity Extraction: From Unstructured Text to DBpedia RDF triples. WoLE@ ISWC.

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

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

[49]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.

Table des abréviations.

DBpedia	Data base pedia
MENA	Moyen-Orient et Afrique du Nord
RDF	Ressource description framework
RDFS	Ressource description framework schema
SPARQL	Simple protocole and RDF query langage
OWL	Web Ontology Language
LOD	Linked Open Data
ALESCO	Arab League Educational, Cultural and Scientific Organization
PNL	Programmation neuro-linguistique
SOC	Systèmes d'Organisation des Connaissances
CERN	Centre européen de recherche nucléaire
URI	L'Universal Resource Identifier
HTTP	L'HyperText Transfer Protocol
HTML	HyperText Markup Language
XHTML	Extensible HyperText Markup Language
XML	Extensible Markup Language
MathML	Mathematical Markup Language
XSLT	Extensible Stylesheet Language Transformations
XACML	Extensible Access Control Markup Language
SVG	Scalable Vector Graphics
DTD	Description de Type de Document
RelaxNG	Regular Language for XML Next Generation

Table des abréviations.

Xpath	Xml path
XSLT	Extensible Stylesheet Language Transformations
SUO WG	Standard Upper Ontology Working Group
SUMO	Standard Upper Ontology
TOVE	Toronto Virtual Enterprise
JSON	JavaScript Object Notation
POStagging	part-of-speech tagging