AVIS DE SOUTENANCE

THESE DE DOCTORAT

Présentée par

Mr : MUSTAPHA KHALFI

Discipline **:** Informatique

Spécialité : Informatique

Sujet de la thèse : Une nouvelle ressource lexicale riche pour la langue Arabe classique : Conversion digitale du dictionnaire ***al=qamus al=muhit***.

Formation Doctorale : Sciences de l'ingénieur Sciences Physiques, Mathématiques et Informatique.

Thèse présentée et soutenue le samedi 26 décembre 2020 à 16h au Centre de conférences devant le jury composé de :

Nom Prénom	Titre	Etablissement	
Mohammed EL MOHAJIR	PES	Faculté des Sciences Dhar El Mehraz de Fès	Président
Azzeddine MAZROUI	PES	Faculté des Sciences Oujda	Rapporteur
Si Lhoussain AOURAGH	PES	Faculté des Sciences Juridiques Economiques et Sociales Salé	Rapporteur
Ilham CHAKER	PH	Faculté des Sciences et Techniques de Fès	Rapporteur
Azeddine ZAHI	PH	Faculté des Sciences et Techniques de Fès	Examinateur
Ouafae NAHLI Arsalane ZARGHILI Rachid BEN ABBOU	Dr PES PH	Institut des linguistiques computationnelles Pise Italie Faculté des Sciences et Techniques de Fès Faculté des Sciences et Techniques de Fès	Directeurs de thèse

Laboratoire d'accueil : Laboratoire Systèmes Intelligents et Applications.

Etablissement : Faculté des Sciences et Techniques de Fès

Centre d'Etudes Doctorales : Sciences et Techniques de l'Ingénieur

Titre de la thèse : Une nouvelle ressource lexicale riche pour la langue Arabe classique : Conversion digitale du dictionnaire ***al=qamus al=muhit***.

Nom du candidat : Mustapha KHALFI
Spécialité : Informatique

Résumé de la thèse

Actuellement, les grandes ressources lexicales acquièrent une pertinence potentielle élevée pour les systèmes d'information et le besoin de ressources lexicales dans les domaines du Traitement Automatique des Langues Naturelles (TALN), est primordial.

Pour contribuer à répondre à ces besoins pour la langue Arabe qui souffre d'une grave pénurie de ce genre de ressources numériques, nous construisons une ressource lexicale à base du célèbre dictionnaire *al=qāmūs al=muḥīṭ* (AQAM). En utilisant une approche basée sur des règles, nous avons conçu un système qui permet d'extraire des informations morpho-syntaxiques, sémantiques et lexicales du célèbre dictionnaire. Nous avons donc obtenu une version numérisée et structurée d'AQAM, enrichie d'informations explicites morpho-syntaxiques et lexicales.

Dans cette thèse, nous décrivons les étapes suivies pour la conversion d'AQAM en un format lisible par machine, en commençant par la phase de segmentation du texte, puis l'extraction des lemmes et des sens, suivie de la phase d'identification du catégorie grammaticale (Part Of Speech), l'extraction d'informations morpho-syntaxiques et lexicales.

Le dictionnaire est très riche de noms propres, nous avons donc dû faire face à une méthodologie automatique permettant de les identifier et les extraire. La tâche d'extraction des noms propres est effectuée sur la base d'une liste de balises définie constituée de mots-clés utilisés par l'auteur d'AQAM.

De plus, la ressource obtenue est enrichie par des traductions anglaises du lemme et des sens simples qui l'accompagnent à l'aide de deux dictionnaires bilingues Arabe-Anglais.

Ensuite, nous présentons un aperçu d'un alignement d'expérience de la section de la lettre ***bā'*** sur le WordNet de Princeton (PWN) et Suggested Upper Merged Ontology (SUMO).

Cette expérience s'est avérée intéressante car elle a révélé que le mappage d'une ressource lexicale arabe sur une ressource anglaise montre des points communs entre les deux langues, mais elle permet surtout de souligner les non-équivalences entre elles.

Toutes les ressources obtenues sont représentées au format XML et distribuées sous licence gratuite sous la plate-forme CLARIN-IT.

Mots clés : Extraction d'informations, lexique arabe, Al Qamus Al Muhit, dictionnaire lisible par machine, ressource lexicale arabe.