UNE APPROCHE VERS LA COMPREHENSION AUTOMATIQUE DES TEXTES ARABES DESTINEE POUR LES SYSTEMES DE QUESTION-REPONSE

Bakari Dakhli Bakari

République Tunisienne
Ministère de l'Enseignement Supérieur, de la Recherche Scientifique

Université de Sfax
Faculté des Sciences Economiques et de Gestion de Sfax

THESE

Présentée à

La Faculté des Sciences Economiques et de Gestion de Sfax

En vue de l'obtention du

DOCTORAT

Informatique

Par

M^me.^ Wided BAKARI

UNE APPROCHE VERS LA COMPREHENSION AUTOMATIQUE DES TEXTES ARABES DESTINEE POUR LES SYSTEMES DE QUESTION-REPONSE

Soutenue le 15 MARS 2018, devant le jury composé de :

M.	**Rafik BOUAZIZ**	Professeur d'enseignement supérieur à FSEG-Sfax	**Président**
M.	**Mohamed JEMNI**	Professeur d'enseignement supérieur à ENSI-Tunis	**Rapporteur**
M^me.^	**Nadia ESSOUSSI**	Professeur d'enseignement supérieur à ISG-Tunis	**Rapporteur**
M.	**Faiez GARGOURI**	Professeur d'enseignement supérieur à ISIM-Sfax	**Examinateur**
M.	**Mahmoud NEJI**	Maître de conférences à FSEG-Sfax	**Directeur de Thèse**

Année Universitaire 2016 - 2017

DEDICACES

Je dédie cette thèse à…

A mon directeur de thèse: M. Mahmoud NEJI

Qui m'a permis de découvrir le monde de la recherche et m'a donné envie de continuer à travers cette thèse. Nulle dédicace ne serait vous exprimer toute ma reconnaissance.

A mon père Abdelhamid & ma mère Maissa(que Dieu la bénisse)

Aucune dédicace ne saurait être assez éloquente pour exprimer ce que vous méritez pour tous les sacrifices que vous n'avez cessé de me donner depuis ma naissance, durant mon enfance et même à l'âge adulte. Vous m'avez particulièrement encouragé et aidé durant toutes mes années d'études. Ma reconnaissance vous est éternelle pour l'éducation et les principes que vous m'avez inculquée. Que ce travail soit preuve de mon éternelle reconnaissance.

A mon mari: Slouma

Quand je vous ai connu, j'ai trouvé l'homme de ma vie, mon âme sœur et la lumière de mon chemin. L'admiration et l'estime qu'impose votre qualité humaine, m'ont poussé et incité pour mener à terme ce travail. Merci pour votre encouragement et votre soutien.

Que dieu réunisse nos chemins pour un long commun serein et que ce travail soit témoignage de ma reconnaissance et de mon amour sincère et fidèle.

A Faicel, Manel, Tmin, Hanen, Hajer, Chokri, Ranim, Maram, Mohamed Amine, Aya, Alaa

Tout simplement, pour avoir toujours su trouver les mots justes dans les moments les plus difficiles et m'avoir apporté chaque fois le réconfort dont j'avais besoin. Vous m'avez soutenu et veillé à mon succès pendant ces années d'étude loin de vous. J'ai pour vous l'estime et l'admiration qu'imposent vos grandes qualités humaines. Veuillez trouver dans ce modeste travail l'expression de mon affection.

Enfin, merci à tous mes proches et ami(e)s, pour leur soutien et leurs encouragements. Je vous dédie ce travail avec tous mes vœux de bonheur, de santé et de réussite.

Wided…

REMERCIEMENTS

Ce travail n'aurait pas été possible sans le soutien et l'encouragement de mon directeur, mon mari, ma famille, mes ami (e)s et mes cher(e)s. Je suis également redevable à tous ces gens remarquables, et très reconnaissante d'avoir eu l'occasion de travailler avec, et être conseillée par chacun d'eux.

Je voudrais d'abord remercier mon directeur de thèse, ***M. Mahmoud Neji*** *qui ne manque jamais de me surprendre avec l'étendue de ses connaissances, la capacité à capturer l'essentiel dans la recherche et guider ses étudiants vers cette compréhension. Ainsi, pour le soutien apporté lors de la réalisation de cette thèse. Je le suis hautement reconnaissante d'avoir su partager son expérience et ses qualités de recherche. J'espère avoir tiré profit pour mes futures années de recherche. Ses connaissances à la fois théoriques et pratiques, associées à son dynamisme scientifique m'ont impressionnée tout au long de mes années d'étude. Monsieur, j'ai eu le privilège de travailler parmi votre équipe et d'apprécier vos qualités et vos valeurs. Merci aussi pour les nombreuses relectures et corrections jusqu'à la soutenance.*

Cette thèse et mes recherches ont bénéficié grandement de travailler avec ***Professeur. Patrice Bellot.*** *Merci pour le temps qu'il m'a accordé pour guider mes recherches à LSIS Marseille. Sa capacité à voir le portrait global de me tenir ciblée sur la bonne voie, d'élargir mon horizon a joué un rôle dans l'achèvement de ce travail. Merci M. Patrice.*

Mes meilleurs remerciements s'adressent aux ***membres du jury.*** *Vous nous faites l'honneur d'accepter avec une très grande amabilité de siéger parmi notre jury de thèse. Merci à Mr.* ***Momahed JEMNI****, professeur à l'université* ***de Tunis*** *et directeur de l'information et de communication à* ***l'ALESCO****, ainsi qu'à Mme.* ***Nadia ESSOUSSI****, professeur à l'université* ***de Tunis*** *d'avoir pris le temps de rapporter ma thèse et pour les rapports détaillés soulevant des questions très pertinentes. Je remercie également Mr.* ***Rafik BOUAZIZ****, professeur à l'université* ***de Sfax*** *d'avoir accepté de présider ce jury. Enfin, j'adresse mes plus sincères remerciements à Mr.* ***Faiez GARGOURI****, professeur à l'université* ***de Sfax,*** *qui a accepté de consacrer de son temps pour examiner cette thèse.*

Mes sincères remerciements s'adressent à mon mari, ***Slouma****, qui a su me remonter le moral et m'encourager quotidiennement pendant cette période, m'a permis de me détendre lorsque j'en avais besoin et m'a apporté son soutien lorsqu'il fallait avancer et qui m'a poussé à m'accrocher pour finir cette thèse malgré des moments de difficulté. Merci d'être toujours disponible pour me conseiller. Merci pour tes compétences et les discussions très profitables sur l'enseignement de l'informatique et la recherche.*

Je tiens à remercier mes parents pour la confiance qu'ils m'ont faire pendant ces longues années d'étude. Cette thèse n'aurait pas été possible sans le soutien et l'amour de ma petite famille : ma mère, mon père, ***Manel, Faicel, Tmim, Hanen, Hajer, Chokri, Ranim, Maram, Mohamed Amine, Aya, Alaa.*** *Je vous remercie pour l'appui sans équivoque.*

J'exprime de même mes remerciements les plus sincères ***aux membres du laboratoire MIR@CL****. Veuillez accepter mes remerciements les plus distingués et l'expression de ma très haute considération.*

Enfin, il est impossible de citer tous ceux qui je suis redevable à: mes ami(e) : ***Sameh, Mabrouka, Fatma, Wiem, Ameni, et mes collègues,*** *tous qui, directement ou indirectement, ont contribué à ce que ce travail arrive jusqu'à là. Je ne peux trouver les mots justes et sincères pour vous exprimer mon affection et mes pensées, vous êtes pour moi des frères, sœurs et des amis sur qui je peux compter.*

Wided…

Tables des matières

Introduction générale 1
Chapitre 1 :Technologies et formalismes de base des systèmes de question-réponse 1
Introduction…….. 7
1. Concepts généraux de la question-réponse 7
1.1 Qu'est ce qu'un système de question-réponse ? 7
1.2 Qu'est-ce qu'une question ? 8
1.3 Qu'est-ce qu'une réponse ? 9
2. A propos de la question-réponse 9
2.1 Historique de la question-réponse 10
2.2 Domaine d'étude pour un système de question-réponse 12
2.3 Taxonomies des systèmes de question réponse 14
3. Architecture typique d'un système de question-réponse 22
3.1 Analyse des questions 22
3.2 Recherche des documents ou des passages 23
3.3 Extraction des réponses 23
4. La question-réponse pour la recherche d'information précise 24
4.1 Qu'est ce qu'une information précise ? 25
4.2 Un SQR est une extension d'un moteur de recherche 25
4.3 Un SQR est une bonne illustration d'une information précise 27
5. Evaluation des systèmes de question-réponse 29
5.1 Présentation de quelques compagnes d'évaluation 29
5.2 Les métriques de validation 35
Conclusion……. 38
Chapitre 2 : Aperçu des systèmes et des approches adoptés en question-réponse 28
Introduction…….. 40
1. Motivations pour un système de question-réponse 40
2. Aperçu de la question-réponse en d'autres langues que l'arabe 43
2.1 Question-réponse en anglais 44
2.2 Question-réponse en chinois et japonais 48
2.3 Question-réponse en français 49
3. La question-réponse en arabe 53
3.1 Les défis de la langue arabe 54
3.2 Principaux systèmes proposés 55
3.3 Principales approches adoptées 59
4. Analyse performante des systèmes de question-réponse arabes 61
4.1 Avantages et limites des systèmes proposés 65
4.2 Tendances actuelles de la question-réponse en arabe 67
4.3 Positionnement de nos travaux de recherche 68

Conclusion……...69

Chapitre 3 : Construction du corpus AQA-WebCorp et fondements théoriques pour une nouvelle approche.....70

Introduction……..... 71

1. L'analyse des questions en arabe..... 71

1.1 Analyse des questions pour les systèmes de question-réponse arabes 71

1.2 Apports de l'analyse de la question lors de l'extraction de la réponse..... 73

2. La construction d'un corpus..... 74

2.1 Utilisation du Web comme une source de corpus..... 74

2.2 Pourquoi introduire de nouveaux corpus pour l'arabe ?..... 75

3. Méthode proposée pour la construction du corpus AQA-WebCorp 77

3.1 La collecte des questions..... 78

3.2 Présentation du corpus AQA-WebCorp..... 79

3.3 Démarche de la construction 81

3.4 Passages générés..... 83

4. La compréhension automatique de textes..... 84

4.1 Motivation pour la compréhension automatique de textes 84

4.2 Applications de la compréhension automatique de textes 85

4.3 Une représentation sémantique pour la compréhension automatique de textes 86

4.4 Une interprétation logique pour la compréhension automatique de textes 91

5. La reconnaissance d'implications textuelles 95

5.1 Motivation..... 96

5.2 Applications de l'implication textuelle 97

5.3 Implication textuelle en question-réponse..... 98

5.4 Traitement logique de l'implication textuelle..... 101

Conclusion……..... 102

Chapitre 4: Une nouvelle approche sémantique et logique pour la question-réponse arabe 99

Introduction……..... 104

1. Motivations 104

2. Démarches de l'approche proposée 105

2.1 Analyse des questions..... 106

2.1.1 Prétraitement des questions..... 107

2.1.2 Transformation des questions..... 107

2.1.3 Traitement linguistique des questions..... 108

2.2 Recherche des documents 110

2.2.1 Identification des documents pertinents..... 110

2.2.2 Sélection des passages pertinents 111

2.3 Analyse des passages..... 112

2.3.1 Nettoyage des passages..... 112

2.3.2 Normalisation des passages..... 112

2.3.3 Segmentation des passages 113
2.3.4 Traitement linguistique des passages 114
2.4 Représentation logique 115
2.4.1 Construction d'une représentation sémantique avec les graphes conceptuels 116
2.4.2 Raisonnement logique à l'aide des graphes conceptuels 127
2.4.3 Transformation des graphes conceptuels en des représentations logiques 129
2.4.4 Détermination de l'implication textuelle 131
2.5 Recherche de la réponse précise 140
2.5.1 Extraction et pondération des réponses candidates 140
2.5.2 Sélection de la réponse précise 141
Conclusion 142

Chapitre 5 : Développement et évaluation d'un système de question-réponse pour la langue arabe 136
Introduction 144
1. Description générale 144
1.1 Présentation de NArQAS 145
1.2 Outils, ressources et techniques utilisés 146
2. Conception et implémentation de NArQAS 150
2.1 Architecture de NArQAS 150
2.2 Vue de fonctionnement 152
2.3 Composants de NArQAS : entrées-sorties 153
2.4 Détails de l'implémentation 157
3. Evaluation et résultats obtenus 162
3.1 Ensemble de données pour l'évaluation 163
3.2 Mesures d'évaluation utilisées 163
3.3 Résultats obtenus par NArQAS 164
4. Analyse des résultats expérimentaux 166
4.1 Cas d'erreurs et traitements d'amélioration 166
4.2 Performance de NArQAS en comparaison avec Qwant et Ask.com 171
Conclusion 175

Conclusion générale 177
Publications de l'auteur 181
Bibliographie 183
Annexes… 202
Annexe A : Sorties des différents modules de système NArQAS 202
Annexe B : Règles d'Abouenour 206

Liste des figures

Figure 1.1: Architecture générique d'un système de question-réponse [Ligozat, 2006] 8
Figure 1.2: Intersection de la question-réponse avec différents domaines de recherche 9
Figure 1.3: Réponse extraite par un moteur de recherche 28
Figure 2.4: Evolution de la question-réponse en arabe depuis son apparition 53
Figure 3.5: Liste des corpus arabe dans plusieurs domaines par rapport à la question-réponse..... 76
Figure 3.6: Source des questions utilisées dans notre corpus 78
Figure 3.7: Types de questions 79
Figure 3.8: Processus de recherche de documents et d'extraction de passages pertinents 81
Figure 3.9: Statistiques des catégories de textes utilisés dans notre corpus..... 83
Figure 3.10: Exemple de graphe conceptuel pour «أكمل التلميذ الدرس»..... 87
Figure 4.11: Les étapes de l'approche proposée 105
Figure 4.12: Les étapes d'analyse de la question..... 106
Figure 4.13: Représentation de la question en une forme déclarative 108
Figure 4.14: Liste des entités nommées extraits par ArNER pour la question Q1 109
Figure 4.15: Sorties de Stanford pour la question Q1 109
Figure 4.16: Analyse morphologique de la question Q1 avec Khoja Stemmer 110
Figure 4.17: Exemple d'un fichier XML émis par ArNER 114
Figure 4.18: Sorties de Stanford pour le passage du texte P1..... 115
Figure 4.19: Sortie de Khoja stemmer pour le passage du texte P1 115
Figure 4.20: Etape de l'implication logique en un système de question-réponse Arabe 116
Figure 4.21: Extrait d'un passage P1 de texte 117
Figure 4.22: Liste de termes extraits d'un passage de texte..... 117
Figure 4.23: Extraction de concepts..... 118
Figure 4.24: Schéma descriptif de désambiguïsation avec l'algorithme de Lesk..... 120
Figure 4.25: Concepts associés aux termes de la figure 4.22..... 121
Figure 4.26: Liste de relations retenues du passage P1 en appliquant les règles d'Abouenour 123
Figure 4.27: Construction d'un graphe conceptuel 123
Figure 4.28: Des exemples d'implication entre entités nommées 136
Figure 5.29: Architecture de NArQAS..... 151
Figure 5.30: Schéma de fonctionnement de notre système..... 152
Figure 5.31: Prétraitement et analyse de la question Q-teste 158
Figure 5.32: Recherche et extraction des passages répondant à la question Q-teste 159
Figure 5.33: Passages de texte extraits du Web pour la question Q-teste..... 159
Figure 5.34: Post traitements et analyses des passages 160
Figure 5.35: Graphe conceptuel et représentation logique de la question Q-teste..... 161
Figure 5.36: Graphes conceptuels et représentations logiques des passages de la question Q-teste 161
Figure 5.37: Implication textuelle et extraction de la réponse précise à la question Q-teste 162

Figure 5.38: Performance de NArQAS 165
Figure 5.39: Performance de la moyenne 165
Figure 5.40: Relations retenues en appliquant les règles de [Abouenour, 2014] 167
Figure 5.41: Analyse de dépendance fournie par Stanford pour la phrase Phr1 167
Figure 5.42: Extraction de relations entre concepts en appliquant la règle 1 ajoutée 168
Figure 5.43: Relations entre concepts retenues en appliquant les règles de [Abouenour, 2014] 168
Figure 5.44: Analyse de dépendance fournie par Stanford pour la phrase Phr2 168
Figure 5.45: Extraction de relations entre concepts en appliquant la règle 2 ajoutée 169
Figure 5.46: Cas de non correspondance entre entités 171
Figure 5.47: Les réponses de Qwant pour la question Q-teste 172
Figure 5.48: Les réponses d'Ask.com pour la question Q-teste 173
Figure 5.49: Performance de NArQAS en comparaison avec Ask.com et Qwant 175

Liste des tableaux

Tableau 1.1: Types des systèmes de question-réponse 17
Tableau 1.2: Classification des systèmes de question-réponse proposée par [Mishra & Jain, 2015] 18
Tableau 1.3: Classification des systèmes de question-réponse arabes 21
Tableau 1.4: Un moteur de recherche vs un système de question réponse [Pho, 2012] 27
Tableau 1.5: Réponse retenue par un système de question-réponse 29
Tableau 2.6: Exemple de questions répondues par JAWEB 56
Tableau 2.7: Définitions selon le processus de correspondance et les ressources du Web 57
Tableau 2.8: Comparaison d'ALQASIM avec autres systèmes proposés en 2012 58
Tableau 2.9: Tâches couvertes par les systèmes de question-réponse 62
Tableau 2.10: Etude comparative des systèmes de question-réponse en arabe (1/2) 63
Tableau 2.11: Etude comparative des systèmes de question-réponse en arabe (2/2) 64
Tableau 2.12: Avantages et limites des systèmes proposés 66
Tableau 3.13: Description de l'analyse des questions: enquêtes arabes 72
Tableau 3.14: Tâches d'analyse des questions couvertes par les études arabes 73
Tableau 3.15: Exemples de questions utilisées dans la construction de notre corpus 80
Tableau 4.16: exemple de synsets des termes : « البرج » et « برج» à partir d'AWN 119
Tableau 4.17: Sens choisis par Lesk de la phrase Ph_2 121
Tableau 4.18: principe de la règle 1 de [Abouenour, 2014] 122
Tableau 4.19: Analyse de dépendance fournie par Stanford 123
Tableau 4.20: Représentation d'une phrase verbale avec le formalisme des graphes conceptuels 124
Tableau 4.21: Représentation d'une phrase nominale avec le formalisme des graphes conceptuels 124
Tableau 4.22: Liste de termes, concepts et relations des exemples de questions 125
Tableau 4.23: Graphes conceptuels correspondants aux exemples des questions du tableau 4.22 126
Tableau 4.24: Représentations logiques des graphes conceptuels des questions du tableau 4.23 129
Tableau 4.25: Résultat de la classification des implications entre FOLH-FOLTs 139
Tableau 4.26: Scores attribués aux passages retenus avec implication 141
Tableau 5.27: Résultats des expériences menées par NArQAS 164
Tableau 5.28: Exemples de questions utilisées pour la comparaison 174
Tableau 5.29: Performance de NArQAS en comparaison avaec Ask.com et Qwant 174

Liste des abréviations

TALN	*Traitement Automatique de Langues Naturelles*
NArQAS	*New Arabic Question Answering System*
RA	*Raisonnement Automatique*
RTE	*Reconnaissance d'Implication Textuelle*
CAT	*Compréhension Automatique de Textes*
Φ	*Opérateur Phi*
RI	*Recherche d'Information*
IHM	*Interaction Homme-Machine*
SQR	*Système de Question-Réponse*
AA	*Apprentissage Automatique*
YA	*Yahoo Answers*
NTCIR	*NII Test Collection for IR Systems*
QA4MRE	*question-réponse pour la machine de lecture*
LCC	*Langage Computer Corporation*
AWN	*WordNet arabe*
AQA-WebCorp	*Arabic Question Answering Web Corpus*
REN	*Reconnaissance des Entités Nommées*
GC	*Graphe Conceptuel*
RL	*Représentation Logique*
LC	*Liste de concepts*
LT	*Liste de Termes*
WSD	*Désambiguïsation des sens des mots*
FOLT	*First Order Logic Text*
FOLH	*First Order Logic Hypotesis*
TREC	*Text REtrieval Conference*
CLEF	*Forum Cross-Language Evaluation*
TAC	*Text Analysis Conference*

Introduction générale…

INTRODUCTION GENERALE

Le travail de cette thèse s'inscrit dans un cadre général du traitement automatique de la langue naturelle (TALN). Il se situe dans le contexte de la recherche d'informations précises. Nous nous intéressons plus précisément aux systèmes de question-réponse. En effet, cette thèse a pour but de proposer une nouvelle approche pour la question-réponse arabe. Cette approche sera implémentée en un système de question-réponse pour l'arabe. Ce système intègre des procédures de raisonnement automatique (RA) et des techniques de reconnaissance d'implications textuelles (RTE). Pour le faire, nous nous appuyons sur des représentations sémantiques et logiques de la question et des passages. Ainsi, nous déterminons l'implication entre ces représentations logiques afin de trouver la réponse précise.

(a) La question-réponse

Un système de question-réponse permet de répondre à des questions en extrayant la réponse à partir des documents. Dans cette thèse, nous nous intéressons à réaliser un nouveau système de question-réponse pour l'arabe. Ce système s'intéresse plus particulièrement sur des questions factuelles, il prend en entrée une question sous forme textuelle comme «من إخترع الحاسوب الألي ؟ » et renvoie la réponse «تشارلز بابيج» ainsi que le passage duquel la réponse a été extraite et justifiée «تشارلز بابيج العالم الأول الذي اخترع الحاسوب».

(b) La compréhension automatique de textes

La compréhension automatique de textes est une tâche ardue du traitement automatique de la langue naturelle. Elle se fonde généralement sur l'hypothèse qu'un texte peut être segmenté en unités de textes (mots, phrases ou paragraphes). D'ailleurs, comprendre un texte consiste à trouver une représentation de son contenu permettant à l'utilisateur de vérifier sa cohérence et de le changer. Ainsi, la compréhension d'un texte peut être définie comme un processus cognitif de compréhension des concepts et des relations qui les relient. Ces concepts et relations sont mis en œuvre pour construire une représentation sémantique des textes via les graphes conceptuels. En fait, pour chaque graphe conceptuel, nous proposons d'extraire la représentation logique correspondante en se basant sur le principe de l'opérateur

Phi (Φ) de [Sowa, 1984]. Pour assurer cette représentation, des analyses linguistiques sont prises en compte pour les questions et les passages du texte.

(c) La reconnaissance d'implication textuelle

La reconnaissance d'implications textuelles consiste à décider pour deux fragments de textes si le sens de l'un est impliqué par celui de l'autre [Dagan & Glickman, 2004]. Elle est utile à un système de question-réponse au sens où un texte est une réponse à une question si une représentation de la question est impliquée par une représentation de ce texte. Dans nos travaux, nous déterminons l'implication textuelle entre des paires des représentations logiques de la question (hypothèse) et du passage de texte qui répond à cette question (texte). Pour le faire, il est nécessaire de comprendre précisément le sens du passage et de la question.

De nos jours, la quantité d'informations se multiple de façon exponentielle avec l'évolution de la toile. Avec l'accroissement de ces informations, accéder à l'information précise est alors une tâche ardue. Ainsi, la croissance de cette quantité d'informations rend la recherche d'une telle information une tâche complexe et coûteuse en termes de temps et de précision. Pour trouver cette information, l'utilisateur adopte communément des moteurs de recherche. Ces derniers retournent un ensemble de documents et délèguent à l'utilisateur la tâche de trouver l'information cherchée [Ben Abacha, 2012]. Ceci entraîne l'utilisateur à effectuer des efforts et des tâches supplémentaires. Malgré que les moteurs de recherche actuels donnent des résultats acceptables, ils rencontrent certaines limites, lorsque l'utilisateur cherche une information précise.

Une voie possible pour répondre à ces limites a motivé le développement de nouveaux outils de recherche adaptés, comme les systèmes de question-réponse [Bernard, 2011]. En effet, un système de question-réponse vise à répondre à une question posée par une réponse courte et précise au lieu d'une liste de documents. Ce système est apparu comme une solution alternative pour les moteurs de recherche dont il produit la notion d'information précise. Il permet également d'économiser le temps de réponse pour l'utilisateur. En conséquence, un système de question-réponse varie avec les moteurs de recherche par leurs données d'entrée et les fins de leurs objectifs.

La recherche en question-réponse se développe tous les jours. La recherche dans le domaine de la question-réponse arabe a commencé dans les années 1990. A notre

connaissance, il y a beaucoup de systèmes qui ont été développés pour de nombreuses langues du monde (ex. anglais, français, chinois, japonais, etc.). Cependant, il existe peu de systèmes qui ont été développés pour la langue arabe. Ces systèmes fournissent des réponses sous forme de passages courts, des extraits de collection de documents. Par conséquent, la performance de ces systèmes est limitée par la difficulté de traitement de la langue et le manque considérable d'outils de TALN efficaces qui prennent en considération l'arabe [Alagha & Abu-Taha, 2015], [Al-Khalifa & Al-Wabil, 2007]. De plus, plusieurs approches ont été proposées pour la syntaxe et la morphologie, très peu d'approches qui sont basées sur la sémantique ont été proposées pour cette langue. A notre connaissance, les approches qui sont fondées sur le raisonnement logique et l'implication textuelle sont rares.

Par conséquent, la principale motivation pour proposer une nouvelle approche pour la question-réponse arabe est que cette langue n'a pas été suffisamment étudiée dans le domaine de la question-réponse. Simultanément, en raison des difficultés de l'arabe et du manque de ressources et d'outils, il est si difficile de fournir des systèmes de question-réponse avec une grande précision. Ainsi, les systèmes de question-réponse arabes sont encore peu nombreux comparés à l'anglais et à d'autres langues latines qui ont beaucoup bénéficié de l'avancement dans ce domaine. Même si, la langue arabe est considérée comme l'une des dix premières langues du monde, elle est moins étudiée par les chercheurs en question-réponse. En conséquence, nos travaux de thèse s'inscrivent dans le cadre de proposer une nouvelle approche à la question-réponse arabe qui combine de nouvelles techniques, y compris, la sémantique, la logique, l'implication textuelle, etc. En réalité, le traitement sémantique et /ou logique dans la question-réponse arabe n'a pas encore reçu suffisamment d'attention.

Notre principal objectif est de proposer une nouvelle approche pour l'arabe qui consiste à l'accouplement de la sémantique avec la logique pour générer des représentations sémantiques et logiques de la question et de ses passages et détermine l'implication textuelle entre ces représentations logiques. Cet objectif peut être réalisé à travers :

- La proposition d'une nouvelle approche qui fournit une compréhension approfondie de la question et des textes à partir de laquelle la réponse doit être identifiée. En effet, une telle sorte de compréhension fournit une représentation sémantique et logique des questions et des passages trouvés.

- La collecte des questions factuelles qui attendent comme réponse une entité nommée.
- L'interrogation d'un moteur de recherche pour retrouver des passages de textes susceptibles de contenir des réponses aux questions collectées.
- L'analyse de la question en extrayant les principales caractéristiques (e.g. les mots clés, le type de la réponse attendue, etc.). En effet, toutes les caractéristiques recueillies sont ajoutées aux étapes suivantes d'un système de question-réponse (e.g. la recherche de documents, la représentation logique, etc.).
- La création d'un corpus de paires de questions-textes en interrogeant le moteur de recherche Google. Cette constitution permettra alors d'offrir une meilleure base pour notre expérimentation.
- L'analyse des passages de textes en présentant les traitements linguistiques qui sont mis en œuvre par des techniques de TALN pour assurer leur analyse. Cette analyse augmente la chance de trouver la réponse précise à la question.
- La représentation des questions et des passages de textes au moyen du formalisme des graphes conceptuels [Sowa, 1984] pour pouvoir générer leurs représentations logiques.
- L'intégration de la logique aux systèmes de question-réponse arabes. Cette intégration se manifeste par la transformation d'un graphe conceptuel de chaque phrase de la question et du passage dans une forme logique. La transformation en des formes logiques est utilisée pour déterminer l'implication textuelle entre une question et les passages de texte qui lui répond.
- La proposition d'un algorithme de transformation des graphes conceptuels en formes logiques en se basant sur le principe de l'opérateur Phi (Φ) de [Sowa, 1984].
- L'inclusion de la technique d'implication textuelle aux systèmes de question-réponse arabes. Cette inclusion est mise en œuvre dans nos travaux pour déterminer la relation d'implication entre deux formes logiques de la question et du passage de texte.
- La réalisation d'un système de question-réponse pour l'arabe qui permet de chercher des réponses à partir du Web à des questions factuelles.

- L'extraction de la réponse qui permet de sélectionner une réponse précise parmi d'autres réponses candidates.

En vue d'atteindre l'objectif mentionné précédemment, nous avons eu recours à diverses techniques telles que la recherche d'informations (RI), le TALN, l'extraction d'informations (EI), les techniques d'intelligence artificielle (IA), le raisonnement automatique et les techniques de RTE.

Cette thèse est organisée comme suit : Après cette introduction, nous allons présenter, dans le chapitre 1 intitulé « **Technologies et formalismes de base des systèmes de question-réponse** », les technologies et les formalises de base pour un système de question-réponse. Comme l'indique son nom : « **Aperçu des systèmes et des approches adoptés en question-réponse** », le chapitre 2 dresse un survol de la littérature des systèmes et des approches de question-réponse dans quelques langues telles que l'anglais, le français, le japonais, le chinois et l'arabe. Dans le chapitre 3 nommé « **Construction du corpus AQA-WebCorp et fondements théoriques pour une nouvelle approche** », nous décrivons la méthode proposée pour la construction d'un corpus de questions-textes et nous illustrons les fondements théoriques pour une nouvelle approche en arabe. Le chapitre 4 est intitulé « **Une nouvelle approche sémantique et logique pour la question-réponse arabe**». Dans ce chapitre, nous décrivons une nouvelle approche sémantique et logique pour la question-réponse arabe qui se fonde sur la compréhension automatique de textes pour répondre à une question. Le chapitre 5 intitulé « **Développement et évaluation d'un système de question-réponse pour la langue arabe** », décrit la conception et la réalisation d'un système de question-réponse pour l'arabe. Il présente les résultats expérimentaux obtenus par ce système. Nous clôturons cette thèse par une conclusion générale et quelques perspectives d'amélioration.

CHAPITRE 1 :TECHNOLOGIES ET FORMALISMES DE BASE DES SYSTEMES DE QUESTION-REPONSE

Introduction…. 7

1. Concepts généraux de la question-réponse 7

1.1 Qu'est ce qu'un système de question-réponse ? 7

1.2 Qu'est-ce qu'une question ? 8

1.3 Qu'est-ce qu'une réponse ? 9

2. A propos de la question-réponse 9

2.1 Historique de la question-réponse 10

2.2 Domaine d'étude pour un système de question-réponse 12

2.3 Taxonomies des systèmes de question réponse 14

3. Architecture typique d'un système de question-réponse 22

3.1 Analyse des questions 22

3.2 Recherche des documents ou des passages 23

3.3 Extraction des réponses 23

4. La question-réponse pour la recherche d'information précise 24

4.1 Qu'est ce qu'une information précise ? 25

4.2 Un SQR est une extension d'un moteur de recherche 25

4.3 Un SQR est une bonne illustration d'une information précise 27

5. Evaluation des systèmes de question-réponse 29

5.1 Présentation de quelques compagnes d'évaluation 29

5.2 Les métriques de validation 35

Conclusion…… 38

Introduction

Nous procédons ce chapitre par un aperçu global sur le champ de la question-réponse. Nous présentons également des définitions pour cette discipline, y compris la question-réponse, un système de question-réponse, etc. Nous exposons uniformément quelques technologies de la question-réponse. D'abord, nous précisons l'historique d'apparition de cette technologie. Ensuite, nous exposons un aperçu des systèmes de question-réponse vis-à-vis le domaine traité, ouvert ou restreint. Puis, nous étudions quelques taxonomies présentées dans la littérature de ces systèmes. En effet, dès son apparition, un système de question-réponse est constitué de plusieurs composants plus ou moins indépendants, tels que l'analyse de la question, la recherche de documents, l'extraction de la réponse, etc. A cet égard, nous focalisons sur une architecture typique d'un tel système de question-réponse. De plus, nous montrons, en quelque sorte, la question réponse est considérée comme une bonne illustration de l'information précise. Ainsi, nous précisons l'évaluation des systèmes de question-réponse. Tout va bien en présentant quelques compagnes de validation ainsi que la majorité des métriques de validation utilisées pour les différents systèmes. Nous clôturons ce chapitre par quelques limites et avantages des systèmes de question-réponse.

1. Concepts généraux de la question-réponse

Dans cette section, il est utile de donner des définitions préliminaires de la discipline de question-réponse, y compris la question-réponse, un système de question-réponse, etc. Les définitions mises en œuvre dans ce cadre sont donc représentatives, elles sont utilisées pour la compréhension de concepts de base dont nos travaux de recherche sont focalisés.

1.1 Qu'est ce qu'un système de question-réponse ?

La question-réponse est conçue comme un type particulier de recherche d'information précise. Elle consiste à trouver des réponses courtes et précises à des questions en langage naturel [Wren, 2011], [Bauer & Berleant, 2012]. Selon [Cao et al., 2011], cette discipline est une forme avancée de la recherche d'information. C'est une évolution importante des systèmes de recherche d'informations. La question-réponse est un processus complexe. Il demande d'abord la compréhension d'un besoin d'information exprimé en langue naturelle par une question [Bélanger, 2006]. En effet, il assure la réduction d'un fardeau de multiples documents qui peuvent être assez fastidieux. Simultanément, cette technologie est conçue

pour minimiser le temps de recherche et de navigation et maximiser l'utilité des connaissances scientifiques et de données.

Cette technologie a potentiellement accompagné par une évolution parallèle à celle de la croissance exponentielle et continue de l'information. En particulier, grâce à l'impact crucial de ces informations sur la recherche d'informations et les applications du monde réel, la demande est toujours en croissance pour les systèmes de question-réponse. Ces derniers fournissent aux chercheurs la possibilité de trouver l'information précise [Athenikos & Han, 2010].

Plusieurs définitions ont été proposées à ce jour pour un système de question-réponse. Selon Usunier et ses collègues, un système de question-réponse est un moyen de génération des réponses exactes et pertinentes à des questions formulées en langage naturel [Usunier et al., 2004]. En outre, ce système est capable de répondre à des questions en cherchant la réponse dans un corpus de textes ou sur des sites Internet [Grappy & Grau, 2011]. Dans la plus part des définitions, un système de question-réponse est classiquement constitué d'un ensemble de modules. Ces derniers réalisent respectivement une analyse de la question, une recherche de portions de documents pertinents et une extraction de la réponse.

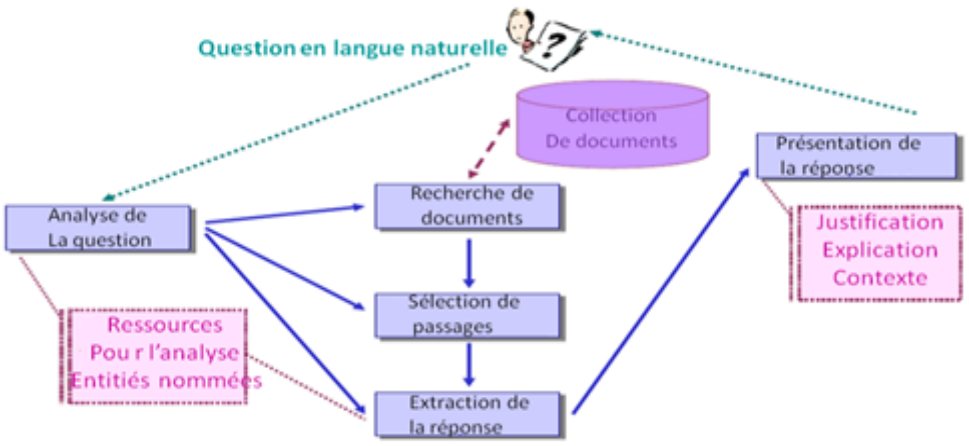


Figure 1.1: Architecture générique d'un système de question-réponse [Ligozat, 2006]

1.2 **Qu'est-ce qu'une question ?**

Une question est une phrase du langage naturel, qui commence habituellement par un mot interrogatif et exprime un besoin d'information de l'utilisateur. Parfois, une question a une forme de construction impérative et commence par un verbe. Dans un tel cas, la demande d'information est appelée déclaration [Kolomiyets & Moens, 2011].

1.3 **Qu'est-ce qu'une réponse ?**

Avec un système de question-réponse, nous devons apporter à l'utilisateur une réponse à la question qu'il a formulée. Or, cette notion qui semble intuitive pose des problèmes de définition. Tout d'abord, la notion de « réponse » qui peut référer à un très court fragment de texte aussi bien qu'a une longue phrase justificative n'est pas clairement définie dans le langage courant. En conséquence, les personnes peuvent répondre d'une manière différente à une même question.

Classiquement, la quasi-totalité des systèmes de question-réponse possèdent des architectures communes mais cela ne signifie pas qu'ils soient similaires. La différence principale entre ces systèmes réside dans l'approche proposée pour chacun d'eux. De surcroît, cette différence se survient dans les techniques et les outils utilisés par ces systèmes. Ainsi, pour quelle langue, un tel système est mis en place.

2. **A propos de la question-réponse**

La question-réponse constitue un champ d'étude majeur en recherche d'informations précises, plus particulièrement, dans le domaine de génération des réponses à des questions posées en langage naturel. En effet, les notions des systèmes de recherche d'informations et des systèmes de question-réponse sont étroitement liées via des concepts qui sont situés à l'intersection de plusieurs domaines, dont notamment le TALN, la recherche d'informations (RI) et l'interface homme-machine (figure 1.2).

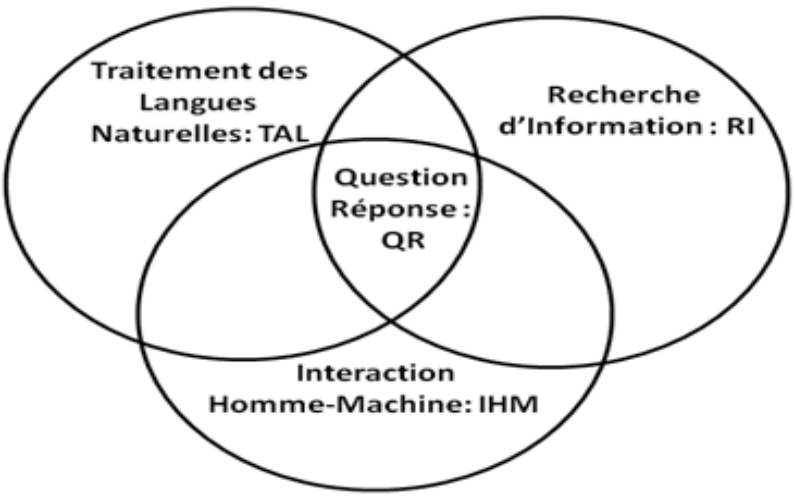


Figure 1.2: Intersection de la question-réponse avec différents domaines de recherche

De nos jours, les usagers qui demandent des informations précises ont besoin d'une vision synthétique et globale des informations afin de guider et d'adapter leur prise de décision. Pour faciliter ce processus, ils utilisent des systèmes de question-réponse. Ces outils permettent aux demandeurs d'informations d'avoir choisi parmi des masses de documents seulement ceux qui peuvent contenir l'information désirée et exacte ainsi de l'extraire.

2.1 **Historique de la question-réponse**

La question-réponse est une technologie qui tente de trouver des réponses à des questions en langue naturelle dans des grandes collections de documents. Un système de question-réponse est apparu comme une solution alternative pour les moteurs de recherche dont il produit la notion d'information précise. Un tel système permet d'économiser beaucoup le temps de réponse pour l'utilisateur. L'objectif principal de tous les systèmes de question-réponse est de récupérer les réponses aux questions plutôt que des documents complets ou des meilleurs passages correspondants.

Dans la littérature de question-réponse, plusieurs tentatives de travaux ont réussi d'illustrer l'historique d'apparition de cette discipline. Depuis 1960, quand ce champ a vu le jour, un ensemble de bases de données liées à la langue naturelle ont été créés comme des systèmes de dialogue, de compréhension du langage, etc. L'idée d'un système de question-réponse est née en 1950 lorsque Turing a présenté une tâche connue comme «un jeu de l'imitation» qui devenait ensuite célèbre connue comme «test de Turing». Depuis que Turing a proposé de considérer la question « Can machine think ? » un être humain peut communiquer avec une machine via une interface (téléscripteur) qui peut être aussi des questions de celui-ci. Turing peut être envisagée comme une machine quand un humain ne pouvait pas faire la différence entre une réponse de la machine et une autre de l'être humain.

Dans ce contexte, plusieurs systèmes de question-réponse ont vu le jour, ces systèmes ont adopté des approches qui sont basées sur le dialogue homme-machine. En effet, ces approches favorisent des transformations des questions posées en langue naturelle en des requêtes d'interrogation des bases de données. A titre d'exemple, nous indiquons le système BASEBALL [Green et al., 1961] et le système LUNAR [Woods, 1973]. En effet, BASEBALL a répondu aux questions d'une base de données structurée des jeux de base-ball et les statistiques. Néanmoins, LUNAR a répondu aux questions relatives à des données géologiques lunaires. Dans la même période, le système AQAS [Mohamed et al ., 1973] a été

proposé pour la langue arabe. Ce dernier est basé sur la connaissance et sélectionne des réponses seulement à partir de données structurées.

En 1978, Lehnert a proposé une approche fondée sur la psychologie à propos de la compréhension du langage afin d'élaborer des systèmes fournissant une interface en langage naturel. Cette approche est validée par le système QUALM [Lehnert, 1977] qui est destiné à la compréhension de textes en domaine ouvert. Après quelques ans, ce genre de travaux porte sur des approches génériques basées sur la détection d'objets, de leurs attributs et des relations des questions, ainsi que sur la traduction des éléments lexicaux en des jetons de chaîne qui mis en œuvre pour décrire les entrées des bases de données [Nguyen & Huong, 2008].

Au cours des années 1980 et 1990, les systèmes de question-réponse, qui sont généralement dédiés pour des domaines restreints sont devenus très populaires. En outre, l'utilisateur est confronté à un certain problème lorsqu'il cherche une réponse. Par ailleurs, l'accès à la base de connaissances est généralement organisé à travers des menus ou des interfaces en langage naturel. Le système lui-même demande interactivement les utilisateurs d'autres questions afin de mieux comprendre son intention. A titre d'exemple nous citons le système de MYCIN [Edward, 1976] qui a été conçu pour offrir une explication des concepts médicaux.

Depuis la fin des années 1990, l'intérêt de la question-réponse a augmenté au sein de la communauté de la recherche, en particulier lorsque la piste de l'introduction de question-réponse a commencé avec TREC-8 en 1999 dans les conférences de recherche de texte [Hirschman et al, 1999]. TREC a eu un impact majeur sur les intérêts en question-réponse et sur le développement des mesures d'évaluation qui comparent les performances de différents systèmes.

La recherche en question-réponse a établi à la fois des systèmes en domaine ouverts ou restreints. Cette piste qui a été instancié par la campagne d'évaluation TREC a lieu régulièrement chaque année depuis 1999 [Voorhees, 2001], [Voorhees, 2004] et [Voorhees & Weischedel, 2000]. Certains systèmes qui sont développés pour un domaine ouvert comme Webopedia [Hovy et al., 2000], Mulder [Kwok et al., 2001] et Answerbus [Zheng, 2002]. Certains d'autres sont conçus pour un domaine restreint, tels que Start [Katz et al., 2002], Naluri [Wong, 2004], WebCoop [Benamara, 2004], ExtrAns [Rinaldi et al., 2004], EpoCare

[Niu & Hirst, 2004] et le système proposé par [Terol et al., 2007]. La plupart des questions abordées par ces systèmes sont des questions factuelles.

En 2011, le potentiel de la question-réponse est mis en œuvre avec le dernier succès de Watson d'IBM sur les jeux de Jeopardy [Wren, 2011]. En effet, Watson analyse les questions pour obtenir ce qui est demandé. En moins de trois secondes, il lance 200 millions de pages en langage naturel qui contiennent sa mémoire pour trouver la bonne réponse et fournit la preuve d'exactitude de la réponse. A part les jeux de télévision, la technologie de question-réponse fournit de telles performances dans d'autres secteurs, tels que la médecine, la météorologie, le voyage, etc.

Les systèmes de question-réponse ont répondu aux questions des utilisateurs après la recherche et le traitement des informations dérivées de plusieurs sources de données, comme le web sémantique [Lopez et al., 2011], [Dwivedi, 2013] et [Suresh kumar & Zayaraz, 2014]. Le format des réponses va également être changé de texte simple au multimédia [Voorhees & Weischedel, 2000].

2.2 **Domaine d'étude pour un système de question-réponse**

Les systèmes de question-réponse sont utilisés par des utilisateurs que veulent connaître des informations en un domaine ouvert ou fermé. En se référant à cette définition, nous distinguons deux grandes classes des systèmes à savoir : systèmes de question-réponse en domaine fermé ou restreint et d'autres en domaine ouvert. Le développement des systèmes de question-réponse en domaine ouvert est considéré comme un élargissement des moteurs de recherche. Souvent, un tel système en domaine ouvert vise à retourner une réponse à une question en langue naturelle. Par conséquent, les questions peuvent concerner tous les sujets. Cette réponse prend la forme de textes courts plutôt qu'une liste de documents jugés pertinents. Dans ce cadre, [Fader et al., 2014] suggèrent que les systèmes de question-réponse ouverts ont besoin de vastes connaissances pour atteindre une couverture élevée.

De leur part, [Ferrucci, 2012] montre que la recherche en domaine ouvert de la question-réponse nécessite des progrès dans plusieurs champs de l'intelligence artificielle (IA) et de l'informatique, y compris, la représentation des connaissances, l'apprentissage automatique (AA), les interfaces homme-machine (IHM), le traitement du langage naturel (TALN), le raisonnement automatisé ainsi la recherche d'informations (RI). Les techniques

qui utilisent toutes ces technologies au même temps pour traiter le langage et la connaissance possèdent un long chemin à suivre avant que les ordinateurs puissent interagir.

Cependant, en un domaine fermé ou restreint la recherche de l'information précise exige un tel système de question-réponse de trouver l'information dans des collections de documents spécifiques à un domaine. Dans ce cadre, les auteurs en [Mishra & Jain, 2015] montrent que le référentiel des patrons des questions est très limité ainsi que les systèmes peuvent obtenir une bonne précision pour répondre aux questions. Alternativement, un domaine fermé pourrait se référer à une situation où seulement un type limité de questions est accepté, telles que les questions qui demandent une descriptive plutôt que des informations de procédure [Mervin, 2013]. Un certain nombre de systèmes ont été conçus pour répondre aux questions en impliquant généralement des domaines spécifiques. Dans ce cadre, les premières approches dans un domaine fermé sont introduites en 1961 avec le système de BASEBALL [Green et al, 1961] et en 1973 avec le système LUNAR [Woods, 1973]. Parmi les autres systèmes les plus connus et qui sont conçus pour un domaine spécifique, nous trouvons WEBCOOP [Benamara, 2004], ce système est dédié au domaine touristique, il utilise une représentation basée sur la logique des données et facilite son interrogation avec des requêtes en langage naturel.

Dans ce contexte, les travaux de [Frank et al., 2007] proposent un système de question-réponse pour les sources de connaissances structurées; [Demner-Fushman & Lin, 2007] présentent leur système pour la question-réponse en médecine; [Ou et al., 2008] propose une approche à base d'ontologie pour la question-réponse dans le contexte des films et cinémas; certaines d'autres études intéressantes ont été présentées dans le cadre du CLEF (Cross-Language Evaluation Forum) et dans l'atelier NTCIR, y compris la réponse aux questions en utilisant Wikipedia; la question-réponse sur la législation européenne [Glöckner & Pelzer, 2010], [Agirre et al., 2010], et la question-réponse avec le raisonnement géographique GikiCLEF [Dornescu, 2010], [Larson, 2009].

En arabe, nous constatons qu'à l'exception de certaines investigations comme [Mohamed et al., 1993] et [Abdelnasser et al., 2014] qui ont réussi relativement à développer les systèmes de question-réponse (AQAS et Al-Bayan) respectivement pour la radiation et le Saint Coran. Le reste des systèmes ont été conçus pour un domaine ouvert, y compris, QARAB [Hammo et al., 2004], ArabiQA [Benajiba et al.,2007], QASAL [Brini et al.,2009],

DefArabicQA [Trigui et al.,2010], AquASys [Bekhti & Al-Harbi, 2013], IDRAAQ [Abouenour et al.,2012], ALQASIM [Ezzeldin et al., 2013], JAWEB [Kurdi et al., 2014], etc.

2.3 **Taxonomies des systèmes de question réponse**

Dans cette section, nous présentons les différentes classifications effectuées pour les systèmes de question-réponse dans différentes langues. Sur la base de la littérature étudiée, nous identifions huit classifications disponibles pour un grand nombre de systèmes de question-réponse. Par ailleurs, nous présentons plusieurs travaux qui favorisent des classifications de ces systèmes. Nous nous limitons de représenter juste les travaux traitant les langues les plus connues telles que l'anglais, le français, l'arabe, etc. De surcroît, nous présentons également une classification proposée pour les différents systèmes en arabe. En termes de connaissance, les études dans cette langue sont présentées en catégories en fonction de nombreux outils et techniques utilisés. Chacune de ces classifications a pris des critères bien déterminés. Dans la suite de cette section, nous citons en détail quelques-unes d'entre elles [Moldovan et al., 2003], [Athenikos & Han, 2010], [Lopez et al., 2011], [Pho, 2012], [Ben-Abacha, 2012],[Gupta & Gupta , 2012], [Mishra & Jain , 2015], etc.

(a) ***Taxonomie proposée par [Moldovan et al., 2003]***

Cette première classification étudie la complexité des questions et la difficulté du processus d'extraction des réponses [Moldovan et al., 2003]. Ce genre de classification favorise cinq classes de systèmes de question-réponse avec croissance de complexité, y compris :

- Classe 1 : Systèmes répondeurs à des questions factuelles.
- Classe 2 : Systèmes favorisent des processus de raisonnement simples.
- Classe 3 : Systèmes extraient des réponses à partir de multiple sources.
- Classe 4 : Systèmes qui proposent un dialogue interactif avec l'utilisateur.
- Classe 5 : Systèmes capables d'effectuer un raisonnement analogique.

(b) Taxonomie proposée par [Athenikos & Han, 2010]

Quelques années après, Athenikos et Han proposent une classification des systèmes en se basant sur deux critères [Athenikos & Han, 2010]. Cette classification nécessite l'appui sur des connaissances sémantiques, elle est composée de trois classes de systèmes :

- SQR sémantiques,
- SQR basés sur les inférences,
- SQR fondés sur des représentations logiques.

Simultanément, dans ce travail les auteurs présentent une autre classification des systèmes de question-réponse en domaine médical en deux classes telles que :

- Systèmes de question-réponse médicaux sémantiques.
- Systèmes de question-réponse médicaux non sémantiques.

(c) Taxonomie proposée par [Lopez et al., 2011]

Une troisième classification proposée par Lopez et ses collègues. En outre, ces auteurs prennent en considération les sources des réponses et les entrées/sorties des systèmes de question-réponse comme un critère de classification [Lopez et al., 2011]. Plus précisément, en s'appuyant sur les ontologies ou sur les ressources de réponses extraites, ces auteurs présentent un travail illustrant un état de l'art sur la catégorisation de ces systèmes. A cet égard, les auteurs favorisent deux types de classification : Le premier est basé sur les ontologies, ce genre de classification engendre l'existence de trois classes de systèmes, à savoir :

- Interfaces en langage naturel pour les bases de données.
- Questions-réponses à partir de documents textuels.
- Questions-réponses avec des données/textes/langages propriétaires.

Le deuxième type définit une classification suivant les sources des réponses. Il favorise trois classes de SQR, y compris :

- Des systèmes dont les sources sont des bases de données structurées.
- Des systèmes dont les sources sont des textes non structurés.
- Des systèmes dont les sources sont des bases de connaissances sémantiques précompilées.

(d) ***Taxonomie proposée par [Pho, 2012]***

Selon une classification présentée par Pho, les systèmes de question-réponse sont basés sur des méthodes de génération des réponses. Ce genre de classification renferme trois types des systèmes de question-réponse: une première classe de systèmes qui sont fondés sur

les patrons [Wyse & Piwek, 2009], la deuxième repose sur la reformulation de la réponse [Kalady et al., 2010]. L'intégration de ces deux genres de méthodes présente une troisième classe de systèmes. Dans ce contexte Pho développe un système de génération des énoncés en langage naturel, appelé SYNGE-A (System for Natural Langage GEneration of Answers) [Pho, 2012]. Ce système prend en entrée un couple de question-réponse en utilisant des ressources externes, qui lui rend générique, paramétrable et adaptable à tous les systèmes de question-réponse, à savoir : des patrons, des lexiques et des règles grammaticales. L'usage de la troisième catégorie des systèmes favorise la particularité du système SYNGE-A par rapport aux autres systèmes : une génération fondée sur les patrons, ainsi sur les règles grammaticales. En outre, quand la question posée par l'utilisateur a été analysée à l'avance, une étape d'annotation est mise en évidence. La génération des réponses se fonde sur les patrons. Ce qui diffère quand la génération est fondée sur les règles grammaticales dont une analyse syntaxique de la question permet d'obtenir un arbre syntaxique.

(e) Taxonomie proposée par [Ben-Abacha, 2012]

Les travaux de [Ben-Abacha, 2009], [Ben-Abacha, 2012] s'appuient sur l'analyse des questions traitées pour produire deux groupes principaux de systèmes: des systèmes à base des approches surfaciques-syntaxiques et des systèmes à base d'approches profondes-sémantiques.

La première catégorie de ces systèmes ne force pas l'analyse sémantique des questions traitées. L'extraction d'une réponse parvient par la recherche des passages de textes qui contiennent cette réponse. Dans ce cadre, une étape d'indexation est effectuée. Cette étape note les expressions qui présentent le mieux document dont la méthode utilisée pour assurer cette indexation est de rendre un document en un « sac de mots ». L'indexation s'appuie sur des traitements linguistiques de documents (une analyse morphologique et/ou syntaxique) pour assurer son enrichissement. Dans cette catégorie, l'utilisation des techniques de traitement des langues (TALN) pour l'extraction des réponses ne se plie pas l'analyse sémantique des questions et des documents.

Néanmoins, la deuxième catégorie exige obligatoirement une analyse de la question traitée ainsi que les documents qui peuvent la répondre. Cette catégorie représente formellement le sens de la question. En se basant sur cette catégorie, divers travaux ont été trouvés [Niu et al., 2003], [Rinaldi et al., 2004]. Plus précisément, Niu et ses collègues

élaborent les enjeux généraux des technologies pour les question-réponse en médecine. En outre, ces auteurs ont utilisé le format PICO (Patient/Problem, Intervention, Comparison, Outcome) pour identifier des rôles sémantiques dans la question et les textes qui seront utilisés dans l'étape d'extraction des réponses. Alors que, le travail de [Rinaldi et al., 2004] favorise des représentations logiques des questions et des documents.

(f) Taxonomie proposée par [Gupta & Gupta, 2012]

En se basant sur les méthodes utilisées, les auteurs exposent une autre classification des systèmes de question-réponse en deux grandes catégories. La première regroupe les systèmes qui subissent des méthodes de traitement du langage naturel et de recherche d'informations. Les tâches principales effectuées dans cette catégorie sont le marquage des entités nommées, le traitement de la syntaxe, etc. La deuxième catégorie rassemble les systèmes qui exercent un raisonnement avec le langage naturel. Ces deux taxonomies sont considérées très importantes, elles ont comparé les caractéristiques de dimensions différentes, telles que les techniques utilisées, les questions traitées, etc. Le tableau 1.1 donne des détails de cette comparaison.

Tableau 1.1: Types des systèmes de question-réponse

Dimensions	Systèmes basés sur les techniques de TAL et de RI.	Systèmes basés sur le raisonnement avec TAL
Technique	Le traitement de la syntaxe, le marquage de l'entité nommée et la recherche d'information	Analyse sémantique ou raisonnement élevé
Source de données	Documents de textes libres	Base de connaissances
Domaine	Domaine indépendant	Domaine orientée
réponses	Des extraits récupérés	Des réponses synthétisées
Questions traitées	Généralement des questions de types Wh	Plus que des questions de types Wh
Evaluations	Utilise une recherche d'information existante	Non mentionné

A cet égard, ces deux catégories proposent quatre classes de systèmes de question-réponse, à savoir:

- Systèmes de question-réponse basés sur le web.

- Systèmes de question-réponse basés sur la recherche d'information / extraction d'information.
- Systèmes de question-réponse en domaine restreint.
- Systèmes de question-réponse basés sur des règles.

(g) Taxonomie proposée par [Mishra & Jain, 2015]

L'étude de la catégorisation des systèmes en d'autres langues que l'arabe s'achève par l'interrogation de [Mishra & Jain, 2015]. Dans leur travail, ces auteurs proposent, explicitement, une catégorisation des systèmes de question-réponse en huit taxonomies. D'abord, Mishra et Jain discutent les détails leur classification. Ensuite, ils donnent une description générale pour chaque groupe ainsi que pour ses classes. Enfin, ils discutent les avantages et les inconvénients des systèmes pour chaque classe. Dans ce cadre, les auteurs catégorisent les systèmes sur la base de différents critères, tels que les types de questions traitées, les types de sources de données consultées, les types de traitement effectués sur les questions et les sources de données, les types de modèle de récupération, les formulaires de réponses générées et les caractéristiques des sources de données. Cette classification favorise huit groupes de systèmes. Chaque groupe contient plus qu'une classe. Le tableau 1.2 affiche les groupes des systèmes. Il montre également les critères étudiés pour chaque classification ainsi les classes pour chaque groupe.

Tableau 1.2: Classification des systèmes de question-réponse proposée par [Mishra & Jain, 2015]

Groupe	Critères	Classes
Groupe 1	Domaine d'application	Domaine restreint
		Domaine ouvert
Groupe 2	Types de questions posées	Questions factuelles
		Question Liste
		Questions hypothétiques
		Questions de confirmation
		Questions causales
Groupe 3	Types d'analyses effectuées aux questions	Analyse morphologique
		Analyse syntaxique
		Analyse sémantique
		Analyse pragmatique
		Analyse du type de réponse attendu
		Reconnaissance de l'objet des questions
Groupe 4	Types de sources de données	Sources de données structurées
		Sources de données semi structurées
		Sources de données non structurées

		Web sémantique
Groupe 5	Types de fonctions correspondantes utilisées dans les différents modèles de récupération	Définir des modèles théoriques
		Le modèle algébrique
		Les modèles de probabilité.
		Les modèles à base de fonctionnalités
Groupe 6	Caractéristique des sources de données	La taille de la Source
		La langue
Groupe 7	Techniques utilisés	Des techniques de fouille de données
		Des techniques de recherche d'information
		Des techniques de compréhension du langage naturel
		Extraction des connaissances et découverte des techniques
Groupe 8	Formes de réponses générées par les systèmes	Texte extrait
		Extraits ou d'autres multimédias
		Réponse générée

(h) ***Classification des systèmes arabes***

Selon les classifications qui ont été préalablement mentionnées, nous pouvons noter qu'elles sont mises en place pour certaines langues telles que l'anglais, le français, le japonais et le chinois, etc. Cependant, dans cette littérature, nous trouvons une classification pour les systèmes de question-réponse pour l'arabe [Al Chalabi, 2015]. En l'occurrence, ces systèmes sont catégorisés en quatre classes. D'ailleurs, cette classification s'est effectuée à la base de plusieurs outils et techniques qui sont extensivement utilisés par chaque système.

- Systèmes basés sur l'interrogation des bases de données.

Les systèmes adoptés à cette catégorie introduisent une approche fondée sur le dialogue Homme-Machine. En effet, ces systèmes transforment la question en une requête et interrogent des bases de données afin de sélectionner la réponse. AQAS [Mohamed et al, 1973] est considéré parmi les premiers systèmes qui sont apparus des les années 60, il cherche des réponses à partir des bases de données structurées. En revanche, QARAB [Hammo et al., 2004] cherche des réponses à des questions à partir de documents non structurés extraits à partir du journal Al-RAYA.

- Systèmes basés sur les techniques de TAL et de RI.

Cette catégorie illustre les principaux systèmes qui reposent extensivement sur des techniques de TALN et de recherche d'information pour de trouver la réponse précise. En effet, la quasi-totalité des systèmes traitent un type particulier de questions, à savoir, la

question factuelle et reposent sur des approches morpho-syntaxiques. Par exemple, ArabiQA [Benajiba et al., 2007] recourt les techniques de reconnaissance des entités nommées ; QASAL [Brini et al., 2009] repose sur la plateforme Nooj pour extraire la réponse à partir d'un livre d'éducation. Par contre, AQUASYS [Bekhti & Al-Harbi, 2013] permet d'analyser la question et d'extraire la réponse à partir d'un corpus. En plus, JAWEB [Kurdi et al., 2014] a été construit sur la base d'AQUASYS en fournissant une interface utilisateur comme une extension.

- Systèmes basés sur la compréhension automatique de textes.

Une troisième catégorisation repose sur la compréhension automatique de textes pour répondre à des questions. L'extraction de meilleures réponses nécessite un certain type d'inférence et un examen de bases de connaissances acquises précédemment [Banerjee et al., 2013]. Dans ce cadre, la plupart des systèmes favorisant la compréhension automatique d'un texte et utilisent des prétraitements, tels que la résolution d'anaphores, la coréférence, ou la reconnaissance d'entités nommées. En fait, IDRAAQ [Abouenour et al., 2012] participe à la tâche QA4MRE@CLEF. Cette dernière a inclus pour la première fois la langue arabe. D'autre part, ALQASIM proposé par [Ezzeldin et al., 2013] est basé sur la sélection et la validation de la réponse, il répond à questions à choix multiples. Ce système prend en compte la compréhension en lecture des questions.

- Systèmes basés sur la logique et l'inférence.

Cette taxonomie des systèmes de question-réponse s'appuie sur le raisonnement logique et l'inférence textuelle à fin de trouver la réponse précise à une question en langue naturelle. A notre connaissance, ce type d'approches est peu utilisé jusqu'à présent dans la question-réponse arabe. L'usage de la logique et de l'inférence dans ce domaine a fait l'objet des travaux rares, à savoir [NBdour & Gharaibeh, 2013]. Ce dernier travail procure une représentation sémantique contrainte en utilisant un cadre d'unification explicite fondé sur l'expansion de la requête (des synonymes et des antonymes) et des similitudes sémantiques.

En se référant à une analyse détaillée dans le (tableau 1.3), chaque catégorie est présentée avec des exemples de systèmes, des techniques et des outils qu'ils adoptent.

Tableau 1.3: Classification des systèmes de question-réponse arabes

Catégorie	Critères/ techniques utilisés	Exemples de systèmes
Systèmes basés sur l'interrogation des bases de données.	Base de données structurées	AQAS [Mohamed et al., 1993]
	Base de données non structurées	QARAB [Hammo et al., 2004]
Systèmes basés sur les techniques de TAL et de RI.	Reconnaissance des entités nommées	ArabiQA [Benajiba et al., 2007]
	Plateforme NOOJ	QASAL [Brini et al., 2009]
	Patrons lexicaux	DefArabicQA [Trigui et al., 2010]
	Reconnaissance des entités nommées	AquASys [Bekhti & Al-Harbi, 2013]
	Traitement automatique de langue TAL	JAWEB [Kurdi et al., 2014]
	Reconnaissance des entités nommées	AL-Bayan [Abdelansser et al., 2014]
Systèmes basés sur la compréhension automatique de textes.	Distance densité du modèle N-gramme et expansion sémantique	IDRAAQ [Abouenour et al., 2012]
	Sélection et validation de la réponse	ALQASIM [Ezzeldin et al., 2013]
Systèmes basés sur la logique et l'inférence	Recherche d'information RI et TAL et intelligence artificielle IA	Système de [NBdour & Gharaibeh, 2013]

En récapitulant, nous nous sommes intéressés sur des approches fondées sur la sémantique et/ ou la logique. Nous dressons comme objectif de proposer une nouvelle approche sémantique et logique pour la détermination d'implication textuelle. Cette approche combine les techniques de traitement automatique du langage naturel, de recherche d'information, d'intelligence artificielle, de reconnaissance d'implication textuelle. Elle favorise une transformation des textes arabes en des représentations sémantiques et logiques. En l'occurrence, notre approche apporte l'intégration de la sémantique et de la logique pour les systèmes de question-réponse arabes. L'idée est de transformer ces textes en des graphes conceptuels puis en des formes logiques. Puis, nous proposons de déterminer des implications textuelles entre les formes logiques des questions et celles des textes qui leurs répondent.

3. Architecture typique d'un système de question-réponse

Dés son apparition, un système de question-réponse correspond, en général, à une chaîne de traitements collectant trois ou quatre composants qui sont plus ou moins indépendants les uns des autres. Dans ce cadre, plusieurs chercheurs, tels que [Athentikos & Han, 2010], supportent l'idée que les principaux composants pour générer une réponse précise à une question en langue naturelle sont l'analyse de la question, la récupération des documents/des passages et l'extraction de la réponse. Malgré que les techniques se diffèrent d'un système à un autre, une architecture typique emploie généralement une architecture pipeline qui enchaîne ces trois modules. D'ailleurs, chacun de ces composants mérite d'être évaluer de façon intrinsèque, or c'est leur assemblage qui est évalué dans sa globalité.

Dans ce cadre, l'étude de [Bilotti & Nyberg, 2008] suggère que ces modules ne peuvent jamais être totalement découplés parce que l'analyse de la question et d'extraction de la réponse dépendent d'une représentation commune des réponses et peuvent-être aussi un ensemble commun d'outils de traitement de texte. Ainsi, cette dépendance est nécessaire d'activer le mécanisme d'extraction des réponses pour déterminer si ces réponses existent dans le texte récupéré. Dans ce qui suit, nous pouvons décrire brièvement chaque module de cette architecture.

3.1 Analyse des questions

L'analyse des questions est une étape essentielle dans la chaîne de traitement d'un système de question-réponse. Plusieurs travaux soulignent que la tâche de répondre à une question nécessite essentiellement une analyse en profondeur de la question [Embarek, 2008]. Cette analyse extrait les indicateurs clés de la question, à savoir, le type de la réponse attendue, l'objet sur lequel porte la question (focus), les termes qui serviront par la suite à la recherche de documents qui pourraient être des réponses [Zweigenbaum et al., 2008]. Ces caractéristiques pourraient être utiles dans les étapes lors de la recherche des réponses [Rodrigo et al., 2010]. Un autre objectif complémentaire et essentiel pour cette étape vise à reconnaître les entités nommées situées dans les questions et à aborder les relations qui les relient. Dans d'autres cas, l'intérêt de l'analyse de la question est essentiellement focalisé sur des reformulations possibles de la question afin d'augmenter la capacité des autres modules à identifier la réponse [Hasan, 2008]. D'autres travaux portent sur la classification de la question comme la première problématique de recherche en question-réponse [Burger et al.,

2001]. Cette classification fournit des informations sur le genre de la question. En effet, elle implique des traitements de la question pour identifier la catégorie de la réponse que l'utilisateur désire trouver. Cette étape est effectuée en utilisant les informations obtenues lors de la segmentation de la phrase. Cette segmentation implique un système de trouver les noms, les verbes, les adjectifs et les prépositions [Moreale & Vargas-Vera, 2004].

3.2 **Recherche des documents ou des passages**

Le but de recherche des documents et des passages est de récupérer les documents susceptibles de répondre à une question donnée. Ces documents sont les plus liés à l'analyse de la question effectuée dans son étape d'analyse [Weiming & Hu, 2007]. En effet, la recherche de documents est une étape complémentaire de l'analyse de question. Celle-ci sert à inciter les moteurs de recherche traditionnels à récupérer la liste de documents pertinents proprement liés à l'analyse de la question effectuée précédemment ou même des passages restreins contenant la réponse. Par ailleurs, si l'analyse de la question est bien traitée, les documents retournés sont ceux qui contiennent l'information désirée. Nous pouvons ajouter que le rôle des systèmes de question-réponse sert également à éliminer le plutôt possible la redondance de ces documents. Par contre, si l'analyse de la question est mal traitée, le résultat de la recherche sera alors une grande masse documentaire. Ainsi, que cette dernière peut ne pas contenir la réponse souhaitée. Généralement, la recherche de documents varie selon le domaine étudié. En domaine ouvert, la recherche d'une information se fonde principalement sur une collection de documents ou de textes qui couvrent tous les données existants sur le web. En revanche, en domaine fermé, la récupération de documents s'effectue sur une collection limitée de documents [Demner-Fushman & Lin, 2007].

3.3 **Extraction des réponses**

L'extraction de la réponse identifie les réponses candidates de l'ensemble de passages pertinents et extrait la réponse la plus probable pour répondre à la question de l'utilisateur [Mervin, 2013]. Il est à noter que l'objectif final d'un tel système de question-réponse est d'extraire et présenter les réponses aux questions posées en langue naturelle. Ces réponses sont les phrases candidates disponibles après les procédés antérieurs (analyse de la question, recherche de documents et ou des passages). En effet, les phrases pertinentes, qui ont été potentiellement utiles pour répondre à une question, ont été disponibles à l'étape de sélection de passages [Weiming & Hu, 2007]. L'étape d'extraction de la réponse représente la tâche

finalisant le processus de traitement d'un système de question-réponse. En outre, la réponse sélectionnée est une parmi plusieurs réponses qui ont été potentiellement disponibles à l'étape de sélection de passages. A ce stade, extraire une réponse précise sert également à examiner la liste de passages retenus après avoir réaliser la phase de sélection de passages. Puis, elle permet de sélectionner la phrase la plus appropriée à la question formulée. Pour certains cas, la réponse retournée est une réponse unique courte ou un extrait d'un document contenant la bonne réponse avec son contexte [Embarek, 2008].

Il convient de noter qu'il existe une autre catégorie de systèmes de question-réponse qui fournissent des composants supplémentaires, à savoir, la justification de la réponse [Nyberg et al., 2003], l'extension des requêtes en utilisant les ressources externes (e.g. le Web) [Abouenour et al., 2012], [Ganesh & Varma, 2009]. En effet, la justification de la réponse prend la réponse produite par le système et tente de la vérifier à l'aide des ressources, telles que le Web, elle emploie des sources de connaissances ou des bases de données externes afin de générer la réponse et chercher de nouveau pour trouver les bons documents. L'extension des requêtes est souvent effectuée puisque les questions peuvent être très courtes. Donc, prendre seulement les mots clés de la question ne permet pas de produire suffisamment d'informations contextuelles pour une récupération efficace.

4. **La question-réponse pour la recherche d'information précise**

La recherche d'information se divise en deux branches: la recherche d'information classique, qui consiste à rechercher des documents en sebasant sur des mots-clés fournis par l'utilisateur et la question-réponse qui effectue des recherches à partir de questions formulées en langue naturelle. La deuxième branche apporte la notion que l'information recherchée soit précise. En effet, au lieu de fournir une liste de documents à une question donnée, il semble mieux de la répondre d'une manière précise. En outre, un système de question-réponse fournit une réponse précise en utilisant des procédures d'analyse et de raisonnement pour mieux comprendre la question et mieux formuler la requête pour augmenter la pertinence d'un tel système. Simultanément, ce système est une perspective d'apport linguistique à la recherche d'information puisque qu'il donne accès à une information précise. Pour cela, la recherche d'information précise est considérée comme une discipline émergente de la question-réponse. Dans le même cadre, [Zweigenbaum et al., 2008] soulignent que les systèmes de question-réponse sont en réalité une évolution des systèmes de recherche d'information. Ils utilisent

des méthodes essentiellement numériques. Selon ces auteurs, la recherche d'une information précise c'est « la question-réponse ».

4.1 **Qu'est ce qu'une information précise ?**

La notion d'une information précise prend ses origines avec la question-réponse. Dans ce cadre, pour minimiser l'effort de l'utilisateur, il convient de lui permettre d'exprimer son besoin sous une forme plus naturelle, comme une question, et de fournir une réponse précise plutôt qu'une liste de documents. Le domaine de la recherche d'information s'adresse à cette problématique en proposant à l'utilisateur des documents qui répondent au mieux à son besoin d'informations. Ce besoin est transformé en requête, dont le contenu est comparé à celui des documents afin d'évaluer leur pertinence. Ainsi, il n'est pas possible de préciser si l'on cherche à s'informer sur un sujet et à obtenir des documents concernant ce sujet, ou si l'on recherche une information précise comme la date d'un événement, la signification d'un acronyme ou l'évolution d'un phénomène au cours du temps (ce que nous appellerons besoin d'information précis).

4.2 **Un SQR est une extension d'un moteur de recherche**

Les systèmes de question-réponse sont considérés comme une extension des moteurs de recherche d'information. Dans ces derniers, un utilisateur est en mesure de rechercher des informations en utilisant un ensemble de mots clés. Le résultat de cette recherche est un ensemble de documents ou des liens vers les documents que l'utilisateur doit les parcourir pour trouver l'information précise. En revanche, les systèmes de question-réponse consistent à fournir des réponses courtes et pertinentes qui peuvent être textuelles ou parlées. Par exemple, la recherche des principaux acteurs jouant dans le film "Titanic", réalisé par James Cameron, une question possible à un système de question réponse serait: Qui a joué les principaux rôles du film Titanic réalisé par James Cameron?. Comme réponse, le système peut donner: Leonardo DiCaprio et Kate Winslet [Grappy & Grau, 2010].

Les moteurs de recherche renvoient tous les documents où figurent les mots clés de la question. Dans ce contexte, c'est l'utilisateur qui doit explorer ces documents afin de trouver la réponse [Séjourné, 2009]. Egalement, la pertinence de ces documents retournés est mesurée en suivant l'existence des mots dans les documents, la quantité, la visibilité des documents dans la collection, les caractéristiques saillantes, etc.

De son côté, [Pho, 2012] montre que les systèmes de question-réponse se divergent aux moteurs de recherche actuels. Ces divergences peuvent être appréhendées selon trois grands points de vue. Premièrement, les moteurs de recherche répondent à des requêtes des utilisateurs, par contre les systèmes de question-réponse favorisent des réponses à des questions posées en langage naturel. Simultanément, avec les moteurs de recherche l'utilisateur saisie un ensemble de mots clés. Cependant, en utilisant les systèmes de question-réponse, l'utilisateur a la volonté de traiter une question. Une autre différence se trouve également au niveau du résultat retourné par ces deux technologies. En effet, les moteurs de recherche favorisent un ensemble de passages contenant les mots clés de la requête traitée. Alors que les systèmes de question-réponse retournent une réponse précise à la question ou offrent des passages de documents contenant la réponse exacte. Ces détails sont bien affichés dans le tableau 1.4 qui récapitule les principaux points de divergence entre ces deux technologies de recherche d'information.

Quoique que les systèmes de question-réponse et les moteurs de recherche (e.x. Yahoo![1], Google[2], etc.) se ressemblent dont chacun d'eux assurent une interface et répondent aux exigences des utilisateurs en information [Razmara, 2008], [Pho, 2012], il se trouve toujours des différences entre ces deux paradigmes de recherche. En effet, celle la plus distinguable est que les moteurs de recherche s'appuient toujours sur des mots clés afin d'obtenir tous les documents désirés. Ainsi, ils extraient les informations selon un thème général [El-Ayari, 2007]. Ce qui engendre l'utilisateur de balayer un nombre souvent élevé de documents pour trouver la réponse la plus adéquate à sa requête [Ligozat, 2003]. Néanmoins, lors de l'utilisation des systèmes de question-réponse, les utilisateurs traitent des questions en langage naturel. Ces paradigmes se situent à l'intersection de plusieurs domaines, à savoir : l'apprentissage automatique, la recherche d'informations RI, le traitement automatique de la langue naturelle TALN et la représentation des connaissances [Grau & Chevallet, 2008] afin de fournir aux utilisateurs des réponses directes et précises à la place des liens vers des pages ou un nombre important de documents [Athenikos & Han, 2010].

[1] http://www.yahoo.com/
[2] (http://www.google.com)

Tableau 1.4: Un moteur de recherche vs un système de question réponse [Pho, 2012]

Critères de comparaison	Moteur de recherche	Système de question- réponse
Tâche	Répondre aux demandes des utilisateurs.	Répondre à des questions posées en langage naturel.
Entrée	Ensemble de mots-clés.	Question en langue naturelle.
Résultats	Pages contenant les mots-clés de la requête.	Réponse précise (des entités nommées). Réponse courte: phrases. Réponse longue: phrases + justification.

4.3 **Un SQR est une bonne illustration d'une information précise**

Dans les années 90, avec l'apparence de l'internet et l'élargissement au grand public de son utilisation, le besoin d'outils de recherche d'information est devenu crucial. Par conséquent, plusieurs moteurs de recherche on été abordés. Ces moteurs présentent un certain nombre de contraintes pour l'utilisateur, tant au niveau des entrées, dont la syntaxe est soumise à un format spécifique, que des sorties, qui laissent à la charge de l'utilisateur de parcourir en longueur un nombre souvent élevé de documents. Pour résoudre ce problème, les systèmes de recherche d'information ont vu apparaître de nouveaux types d'outils appelés des systèmes de question-réponse. C'est la prise en compte du besoin d'information précise de l'utilisateur qui a soulevé l'émergence de ces systèmes. En effet, ces systèmes tentent de fournir des réponses pertinentes, justifiées pour répondre à ces besoins. Plus précisément, un tel système de question-réponse tente de dépasser la recherche documentaire.

Ainsi, la masse de documents électroniques disponible rend théoriquement possible la recherche d'une information précise, mais pratiquement impossible sans outil de recherche adapté [Ligozat, 2006]. Si les moteurs de recherche actuels sont efficaces afin de trouver des documents correspondants à un besoin d'information large, ils sont moins efficaces pour répondre à un besoin d'information précise. Pour connaître la réponse à une question comme « من اكتشف قانون الجاذبية الأرضية ؟», un utilisateur doit au moins parcourir une liste de documents avant de trouver la réponse, voire adapter sa requête pour optimiser ses chances de trouver un document contenant l'information.

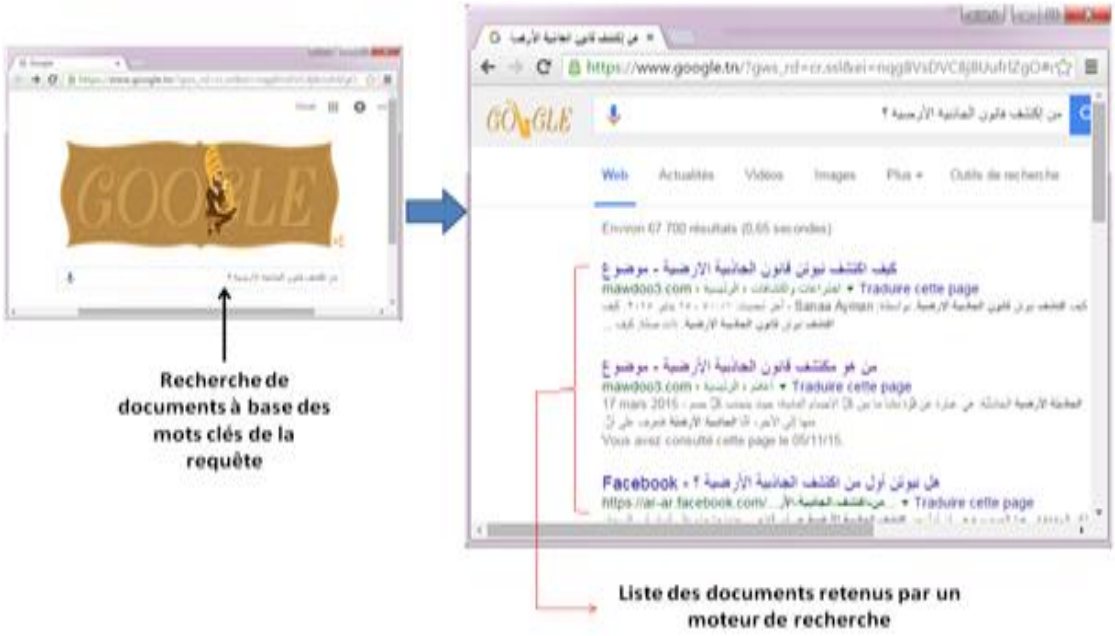


Figure 1.3: Réponse extraite par un moteur de recherche

En répondant à la question au dessus, la figure 1.3 propose une interface alternative pour une recherche simple via le moteur de recherche « Google ». Elle permet de spécifier quatre mots-clés, 'الأرضية', 'الجاذبية', 'قانون', 'اكتشف', etc. Cette interface permet de retourner aux utilisateurs des résultats pour cette requête. Ces résultats pointent vers un grand nombre de liens. Cependant, l'utilisateur sera déstabilisé par le nombre de liens qui croît proportionnellement au nombre de documents.

De sa part, [Ligozat, 2006] a souligné que les systèmes de question-réponse tentent de répondre à de tels besoins. La requête par des mots-clefs d'un moteur de recherche classique est remplacée par une question et la sortie est constituée de la réponse précise à la question, au lieu d'une liste de documents à parcourir. Le domaine de la question-réponse tire ainsi parti des possibilités actuelles en recherche d'information (RI), mais lui apporte une interaction facilitée avec l'utilisateur, grâce à la manipulation et la compréhension du langage naturel.

Ainsi, un système de question-réponse renouvelle le champ de la recherche d'information précise. Il favorise le passage d'un paradigme requête-documents vers un paradigme question-réponse dont les questions sont posées en langue naturelle. Plus précisément, l'utilisateur passe d'une recherche documentaire à une recherche d'informations précises. Ce système décrit, également, un atout majeur pour un moteur de recherche dont les résultats consistent en quelques mots contenant la réponse.

Tableau 1.5: Réponse retenue par un système de question-réponse

Question : من اكتشف قانون الجاذبية الأرضية ؟
Réponse : إسحاق نيوتن

5. Evaluation des systèmes de question-réponse

Nous nous sommes intéressés, dans cette section, à présenter l'évaluation des systèmes de question-réponse. Nous présentons également les métriques de validation. En effet, dans chaque tâche d'évaluation de la question-réponse, tels que TREC, CLEF, NTCIR, EQueR et Quaero, plusieurs systèmes ont été présentés. A cet égard, il est à noter que le but principal d'un système question-réponse était et reste pour toujours de passer de la recherche de documents à la recherche d'information précise par extraction des réponses pertinentes et concises aux questions, dans une grande collection de documents.

De même, l'évaluation des systèmes de question-réponse peut se faire au niveau de la satisfaction de l'utilisateur (applicatif et qualitatif) ou par l'intermédiaire d'une métrique de validation (comparatif et quantitatif). En plus, les campagnes d'évaluation d'un système de question-réponse ont pour but, d'une part, d'estimer les performances de différentes approches. D'autre part, de proposer un certain nombre de questions significatives, avec les catégories les plus fréquentes, telles que factuelle, définition, booléenne, complexe, liste, etc. Dans ce cadre, les systèmes doivent fournir des réponses candidates pour chaque question, celles-ci sont souvent évaluées grâce des métriques de validation.

5.1 Présentation de quelques compagnes d'évaluation

La tâche de question-réponse est incluse dans la majorité de compagnes d'évaluation. La compagne TREC (anglais) introduit cette tâche depuis 1999. Avec des questions en plusieurs langues européennes, la campagne CLEF (multilingue) intègre la tâche QA@clef depuis 2003. Dans la même période (2003), la compagne NTRCIR (japonais) introduit la tâche de question-réponse. Enfin, la campagne nationale Techno langue, EVALDA, comporte le volet EQUER (français) depuis 2004, QUAERO (français, anglais). Dans ces compagnes, les systèmes sont évalués en domaine ouvert et fermé. Ainsi, ces compagnes ont donc l'occasion de tester de nouvelles approches ainsi que de les comparer. Dans ce cas, les

organisateurs de grandes conférences proposent un objectif particulier pour chaque piste spécifique à la question-réponse [Peñas et al., 2012]. Les types des questions analysées sont de types divers (factuelles, définitions, listes, causalités, oui/non, etc.) qui sont les mieux adaptés aux particularités des tâches de chaque occasion de la conférence. Dans ce cadre, les organiseurs choisissent une méthode d'évaluation appropriée qui consiste à déterminer les caractéristiques spécifiques des collections et de sélection des mesures pour évaluer la performance des systèmes participants.

(a) TREC (Text REtrieval Conference)

TREC[3], est l'une des conférences organisées dans le monde annuellement. Elle est destinée à l'environnement de l'évaluation relative de différents systèmes de recherche d'informations à partir de groupes de recherche commerciaux et universitaires [Moldovan et al., 2000]. La tâche principale de cette compagne est la récupération automatique de réponses à des questions factuelles à partir d'un éventail documents [Zhang & Lee, 2002]. Initialement, TREC a été crée en 1992, elle est développée d'une manière significative de sa forme originale. Depuis plus que 17 ans, la piste de question-réponse a été introduite pour la première fois à TREC-8. Cette compagne est coparrainée chaque année par le comité de l'institut national des normes et de la technologie (NIST) et du département de défense américain. Elle fournit plusieurs méthodes à grande échelle pour le processus de validation.

TREC-8 et TREC-9 ont garanti d'une part, que le groupe de documents a une réponse de chaque question [Moldovan et al, 2000]. D'autre part, elles ont nécessité de répondre à des questions factuelles en obtenant un clip de texte qui comprend une réponse à la question. En TREC-2001, le système a la responsabilité de reconnaître la réponse plutôt que de donner une réponse trompeuse [Voorhees, 2001]. La piste de question-réponse pour TREC-2002 a deux tâches : la tâche de la liste et la tâche principale. En outre, la tâche requise par le système était de donner des réponses exactes. Tous les systèmes ont été limités à fournir une réponse pour chaque question et non cinq comme les TRECs précédentes [Soubbotin et al, 2002].

En 2003, la piste de question-réponse fournit un groupe de documents et comprend deux tâches. La tâche principale (comprend trois sortes de questions : les questions de définition, les questions listes et les questions factuelles) et la tâche de passages (répondre à des questions factuelles en obtenant un clip de texte qui comprend une réponse). Dans ce

[3] http://trec.nist.gov/

cadre, les questions nécessitent des réponses courtes fondées sur des faits (500 questions) [Harabagiu et al, 2003]. Dans TREC-2004, des séries de questions utilisées peuvent être définies comme un ensemble d'objectifs, elles étaient à la fois des questions factuelles et listes, elles ne sont pas séparées, toutes sont liées aux objectifs fixés. Plus de membres sont inclus dans la résolution des tâches qui sont à la fois pour les questions listes et définitions [Guo, 2004].

En TREC 2005, une piste connue sous le nom "Robust Retrieval Track" est une tâche qui a utilisé une récupération ad-hoc classique qui se concentre sur l'efficacité d'un sujet unique plutôt que sur la moyenne. En 2006, une autre piste, appelée "Terabyte Track", a été introduite à la question-réponse. Le but de cette piste est de savoir comment la communauté de recherche d'informations est en mesure d'évoluer la recherche d'information traditionnelle pour tester une collection d'évaluation fondée sur les grandes collections de documents.

En TREC-2007, une piste de question-réponse a été conçue pour avoir une étape plus proche de la recherche d'informations plutôt que la recherche de documents. Pour 2008, la conférence, renommée TAC (Text Analysis Conference), s'est tournée plus vers des problèmes d'extraction d'opinion qui sont hors de notre cadre. Dans TREC-2009, une autre piste est utilisée "Million Query Track" ; son but consiste à examiner des hypothèses qui sont générés à partir des sujets avec des décisions incomplètes à la place d'un outil utilisé afin de construire une collection des participants de TRECs traditionnelles. Dans TREC-2010, le but de la piste de question-réponse était d'explorer le comportement de recherche d'information dans la blogosphère.

Dans TREC-2009, une piste entité a été exécutée en tant que système de question-réponse. L'objectif de cette piste était de mettre en œuvre les tâches de l'entité orientée de la recherche dans le Web. Les tâches sont basées sur le retour des objets particuliers au lieu de tout type de document. En TREC-2012, une piste était "piste juridique", son objectif était de développer la technologie de recherche qui reconnaît les besoins des avocats à participer à l'efficacité de découvrir les collections de documents numériques.

Dans TREC-2013, une piste appelée "Crowdsourcing Track" a été annoncée à la découverte de développement des approches fondées sur la foule pour la recherche d'évaluation, ainsi que l'amélioration des systèmes de recherche d'automatisation hybrides. Un exemple des pistes utilisées dans TREC- 2014 "Session Track" a pour objectif de fournir

les ressources nécessaires pour tester les collections qui peuvent aider à évaluer l'utilité du système de recherche d'information en simulant l'interaction de l'utilisateur. Cela peut être fait par une série d'interactions de l'utilisateur et les requêtes au lieu d'une requête qui inclut un seul coup.

En TREC 2015, une nouvelle piste, nommée « Live QA », est menée pour la première fois cette année. Elle est concentrée sur la réponse en temps réel à des questions des utilisateurs. Cette piste a attiré une attention particulière de la communauté de recherche en question-réponse. Les réponses retournées ont été jugées par les éditeurs de TREC sur une échelle Likert à 4 niveaux.

En TREC 2016, la piste de « Live QA » introduite encore une fois. Les questions réelles des utilisateurs, issues du flux des questions les plus récentes sur le site Yahoo Answers (YA) (des questions qui n'ont pas encore été répondues) sont envoyées aux systèmes participants. Ces systèmes procurent une réponse en temps réel. Les réponses retournées sont ensuite jugées par les éditeurs de TREC sur une échelle Likert à 5 niveaux. En se basant sur les résultats de la piste de 2015 et 2016, en 2017 la piste de « Live QA » est diffusée avec une nouvelle sous-tâche portant sur des questions médicales.

(b) CLEF (Cross-Language Evaluation Forum)

La compagne d'évaluation TREC se focalise exclusivement sur l'anglais. En 2003, la piste de question-réponse s'est spécialisée des langues européennes a vu le jour pour la première fois avec un autre cadre d'évaluation européen. Il s'agit du CLEF qui met à disposition de tester différents aspects du développement de systèmes de recherche d'information monolingues et multilingues. Dans cette année, avait lieu la première évaluation des systèmes de question-réponse pour des langues européennes autres que l'anglais. Le but est de promouvoir le développement de systèmes de question-réponse capables, à partir d'une question posée dans une langue source donnée, de retourner une réponse extraite d'une base documentaire dans une langue cible différente. Par conséquent, la question-réponse multilingue est apparue comme une tâche de recherche complémentaire. Dans ce cadre, trois langues étaient mises en places dans la tâche monolingue (néerlandais, italien et espagnol). Pour le faire, CLEF propose un cadre d'évaluation fondé sur le modèle de TREC. Chaque année, plusieurs tâches sont proposées aux participants qui reçoivent un corpus de documents et un ensemble de requêtes. Nous présentons brièvement les tâches

précédentes pour la tâche de question-réponse dans la compagne CLEF. Nous remarquons que chacune de ces tâches se diffère de celles de l'année précédente. En effet, chaque année, il y a une tâche principale qui se compose de systèmes répondant à une série de questions. Cette tâche peut être réalisée soit sur une seule des langues fournies ou avec une autre source et de la langue cible.

L'une des pistes qui a été dirigé à partir de 2003 c'est QA@CLEF: Multilingue Question Answering Track à CLEF, elle est toujours en cours d'apparition chaque année. En effet, depuis 2003, les pistes de CLEF couvrent un champ de compréhension du langage naturel, avec un accent sur le multilinguisme. En 2004, la piste de question-réponse a attiré une attention considérable dans le cadre de CLEF. Elle a impliqué deux tâches différentes et une autre piste: la tâche principale de question-réponse, une tâche pilote espagnol et la piste interactive iCLEF. La piste principale comprenait plus de langues européennes que CLEF 2003 et toutes les combinaisons de langues croisées entre eux ont été exploités pour mettre en place un certain nombre de tâches différentes [Magnini et al., 2004]. En revanche, en 2005, les tâches ont été caractérisées par une continuité de base avec les tâches proposées en CLEF 2004 [Vallin et al., 2005]. Elles étaient notamment inchangées, sauf pour l'ajout de quelques langues qui sont à la fois cibles et sources. Ceci donne aux participants l'occasion d'améliorer les limites actuelles de leurs systèmes.

Dans CLEF 2006, deux tâches ont été ajoutées à la tâche principale des années précédentes [Magnini et al., 2006], WiQA [Jijkoun & De Rijke 2007] et exercices de validation de la réponse (AVE) [Peñas et al. 2006]. Dans la cinquième campagne de question-réponse à CLEF, à la tâche principale, deux autres tâches ont été données, à savoir, AVE qui a continué de succès du pilote de l'an dernier, et QUAST (Question Answering for Speech Transcripts) qui vise à évaluer la tâche de question-réponse dans le discours de transcription [Giampiccolo et., 2007]. De même, dans CLEF 2008, les sous-tâches de l'année précédente ont été données à nouveau, en plus de la tâche principale. La tâche principale est restée inchangée, ce qui permet aux participants d'obtenir plus d'expérience avec les changements de l'an dernier concernant les coréférences dans les questions [Forner et al., 2008].

En 2009 et 2010, trois tâches distinctes ont eu lieu, d'entre elles étant la tâche principale [Peñas et al., 2009]. QAST a été continué, une tâche sur les questions nécessitant un raisonnement géographique, il s'agit de GikiCLEF. Une collection de documents en

parallèle a été utilisée pour la question-réponse multilingue en ResPublicQA. Des questions de raison et d'opinion ont été disponibles. Une nouvelle métrique de validation nommée c @ 1 a été utilisée pour récompenser les systèmes qui réduisent le nombre de questions incorrectement répondues sans affecter la précision des systèmes [Peñas et al., 2010].

Entre 2011-2013: QA4MRE était la nouvelle tâche considérée. Cette tâche exige une connaissance profonde de la signification du texte. Les systèmes utilisent des documents ainsi que la collecte de base pour extraire la réponse. En 2012, même les langues non-européennes, telles que l'arabe ont été introduites dans la piste de question-réponse pour la machine de lecture. Ainsi, l'évaluation et l'analyse comparative des systèmes de question-réponse arabes peuvent être encouragés et soutenus par cette décision. À notre connaissance, la langue arabe a été introduite seulement en TREC 2002, CLEF 2012 et CLEF 2013.

En 2014-2015, trois tâches distinctes ont été mises en considération (QALD (question answering over linked data), BioASQ (Biomedical semantic indexing and question answering), Entrance Exams) dont le point de départ est toujours une question du langage naturel. Cependant, répondre à certaines questions exigent d'interroger les données liées (en particulier si des agrégations ou des inférences logiques sont requises), alors que certaines questions nécessitent des inférences textuelles et des questions sur le texte libre. Enfin, répondre à certaines requêtes peut nécessiter les deux.

(c) Autres compagnes

L'intérêt pour les systèmes de question-réponse a connu un essor important depuis l'introduction de la tâche de question-réponse dans les différentes campagnes d'évaluation. Ceci a commencé par la piste de question-réponse de la campagne TREC (Text REtrieval Conference) du NIST, initiée en 1999. Au delà, d'autres compagnes ont été lancées afin de faciliter l'évaluation des systèmes de question-réponse. La compagne EQueR concerne des recherches sur les systèmes de question-réponse développés pour le français. Cette compagne a été organisée et pilotée par ELDA (http://www.elda.org) [Ayache, et al. 2006]. Les systèmes participants ont mis à disposition des questions factuelles, définitions, listes, ainsi que les questions oui/non ont été introduites pour la première fois. Une autre compagne d'évaluation a eu lieu dans le cadre du projet Quaero [Quaero, 2008], [Quintard, 2009], un grand projet franco-allemand centré sur le contenu numérique, et en particulier l'extraction d'informations, leur analyse et classification, et en général leur exploitation. D'autres évaluations comme

NTCIR (NII Test Collection for IR Systems) qui organise des campagnes d'évaluation translingues entre le japonais et l'anglais ou le chinois et l'anglais introduit la question-réponse.

Par conséquent, l'un des facteurs clés de la réussite dans le champ de la question-réponse est d'organiser des campagnes de validation qui ont aidé les chercheurs dans l'analyse comparative de leurs systèmes selon des mesures standards. La succession de pistes de la question-réponse annuelles, telles que TREC et CLEF a permis l'amélioration de la performance dans une mesure mature pour les langues considérées, notamment, avec le développement des tâches de question-réponse les plus avancés telles que la question-réponse pour la compréhension de lecture (QA4MRE). En raison de l'absence de l'arabe dans les majorités de ces pistes, les systèmes de question-réponse arabes possèdent de nombreux inconvénients en termes de leur processus d'évaluation.

5.2 **Les métriques de validation**

Depuis plus que vingt ans, plusieurs compagnes d'évaluation ont émergé. Ces compagnes ont fixé des normes pour les méthodes d'évaluation. Ces normes sont appelées des métriques de validation telles que (MRR, précision, rappel, F-mesure, etc.). Tout d'abord, nous allons présenter ces différentes métriques fixées dans ces campagnes ainsi que les traitements sur lesquels elles permettent d'insister.

De leur coté [Olvera-Lobo & Gutiérrez-Artacho, 2015], ces auteurs mettent en surbrillance que l'évaluation des systèmes de question-réponse est considérée un domaine de recherche important a besoin de plus d'attentions, particulièrement avec l'apparition du domaine orienté des systèmes de question-réponse basés sur la compréhension du langage naturel et le raisonnement [Sing et al., 2005]. Bien qu'il existe différentes analyses qui se rapportent à l'évaluation des systèmes de question-réponse. A titre d'exemple, nous citons ceux qui sont basés sur l'évaluation des systèmes sur le Web [Radev et al., 2002], [Olvera-Lobo & Gutiérrez-Artacho, 2011] ou dans certains des forums internationaux d'évaluation [Peñas et al., 2012]. Cependant, jusqu'à présent, il n'a eu aucune analyse globale de l'usage des mesures d'évaluation de ces systèmes dans les principaux forums internationaux. De son côté [Abouenour, 2014] a indiqué que l'un des facteurs clés de la réussite dans le domaine de la question-réponse est l'organisation des campagnes d'évaluation. Celle-ci autorise les

chercheurs à effectuer une analyse comparative de leurs systèmes via des mesures standards ; parmi ces mesures nous citons :

(a) Le Rappel :

Le rappel est proportion de documents pertinents trouvés divisé par le nombre total de documents.

$$\text{Rappel} = \frac{\text{Nombre de documents pertinents trouvés}}{\text{Total de documents pertinents trouvés}}$$

(b) La Précision :

La précision est la proportion de documents pertinents parmi les documents trouvés. Elle permet d'estimer la capacité du système à ramener dans les premières positions des documents importants et d'exclure de ces positions des documents non-pertinents.

$$\text{Précision} = \frac{\text{Nombre de documents pertinents trouvés}}{\text{Nombre de documents trouvés}}$$

La Précision et le Rappel sont deux mesures très utilisées en recherche d'information. La précision est une mesure de l'exactitude, le Rappel est alors une mesure de son exhaustivité. Elles peuvent être à la fois combinées dans une moyenne harmonique pondérée appelée F-mesure.

(c) La F-mesure :

La F-Mesure est une évaluation intermédiaire entre la précision et le rappel. Elle représente leur moyenne harmonique pondérée, la formule utilisée pour F est:

$$F = \frac{2 * Precision * Rappel}{Precision + Rappel}$$

(d) La MRR: Moyenne Rang Réciproque

MRR (Mean Reciprocal Rank) est une mesure utilisée dans la question-réponse de TREC. Elle sert à évaluer des processus produisant une liste de réponses possibles à une question. C'est l'inverse multiplicatif de la première réponse exacte et précise. La valeur MMR pour l'expérience est calculée en prenant la moyenne des scores pour toutes les questions [Voorhees, 2001]. En outre, MRR peut être utilisée avec plusieurs réponses

correctes et précises, mais elle ne prend en compte que celle qui est trouvée la première, la plus correcte et fiable. Cette métrique assure également à retourner des réponses correctes appropriées aux premières positions qui possèdent un score de confiance élevé [Barbier, 2009]. Le processus d'évaluation en utilisant MRR se produit comme suit : si une question obtient la première réponse est correcte alors celle-ci reçoit un score de 1, si la réponse correcte est la deuxième, le score sera 1/2, 1/3 si elle est la troisième réponse, et ainsi de suite. Si la réponse n'est pas trouvée, un score de 0 est attribué). La formule pour calculer la MRR est la suivante:

$$\mathrm{MRR} = \frac{1}{Nb - questions} \sum_{i=1}^{\mathrm{Nb}\,questions} \frac{1}{\mathrm{rang}_{-\mathrm{reponse\ i}}}$$

(e) L'exactitude:

L'exactitude est utilisée afin d'évaluer la qualité globale d'un système de question-réponse qui fournit une réponse potentielle. Cette mesure est un nombre compris entre 0 et 1 qui indique la probabilité que ce système fournira la réponse correcte en moyenne. Elle est exprimée comme suit:

$$\mathrm{Exactitude} = \frac{\mathrm{Nombre\ de\ réponses\ correctes}}{\mathrm{nombre\ de\ questions}}$$

(f) La C@1

C @ 1 est une extension de la mesure d'exactitude (la proportion des questions correctement répondues). Elle a été introduite en premier lieu en ResPubliQA [Peñas et al., 2009]. Cette mesure encourage les systèmes pour réduire le nombre de réponses incorrectes tout en conservant le nombre des réponses correctes et en abandonnant des questions sans réponse. Selon [Peñas & Rodrigo, 2011], ne pas répondre a plus de valeur que de répondre de manière incorrecte. Cette mesure a un bon équilibre de la puissance de la discrimination, de la stabilité et de la sensibilité des propriétés. Elle est représentée par la formule suivante:

$$\mathrm{C@1} = \frac{(\mathrm{NR} + \mathrm{NU} * (\mathrm{NR}/\mathrm{N}))}{\mathrm{N}}$$

Avec:

- NR: est le nombre de questions répondues correctement.

- N: est le nombre total de questions.
- NU: est le nombre de questions sans réponse.

Conclusion

Nous avons dressé, dans ce chapitre, un aperçu sur la question-réponse. D'abord, nous définissons quelques concepts de base. Ensuite, nous exposons un historique de la question-réponse. Puis, nous distinguons quelques domaines d'étude pour un système de question-réponse. Nous exposons également quelques classifications proposées de ces systèmes. En effet, nous présentons le fonctionnement des systèmes de question-réponse et détaillons les différents modules intervenant dans la chaîne de traitement, allant de l'analyse de la question jusqu'à l'élaboration de la réponse souhaitée en passant par la recherche des documents. En outre, nous présentons un système de question-réponse comme une extension d'un moteur de recherche. Nous terminons ce chapitre par décrire les différentes compagnes de validation ainsi que certaines mesures étudiées pour la validation.

Dans le chapitre suivant, nous allons présenter un survol de la littérature de différentes approches proposées pour plusieurs langues connues du monde. Nous présenterons aussi quelques systèmes existants dans le but de montrer les différentes techniques utilisées. De plus, nous allaons fournir une analyse performante des études proposées pour la langue arabe. Enfin, nous allons terminer le chapitre en exposant les avantages et les inconvénients des systèmes de question-réponse en arabe et en présentant le positionnement de nos travaux.

CHAPITRE 2 : APERÇU DES SYSTEMES ET DES APPROCHES ADOPTES EN QUESTION-REPONSE

Introduction ……..... 40

1. Motivations pour un système de question-réponse..... 40

2. Aperçu de la question-réponse en d'autres langues que l'arabe..... 43

2.1 Question-réponse en anglais..... 44

2.2 Question-réponse en chinois et japonais..... 48

2.3 Question-réponse en français..... 49

3. La question-réponse en arabe..... 53

3.1 Les défis de la langue arabe..... 54

3.2 Principaux systèmes proposés..... 55

3.3 Principales approches adoptées..... 59

4. Analyse performante des systèmes de question-réponse arabes..... 61

4.1 Avantages et limites des systèmes proposés..... 65

4.2 Tendances actuelles de la question-réponse en arabe..... 67

4.3 Positionnement de nos travaux de recherche..... 68

Conclusion……..... 69

Introduction

Ce chapitre illustre un aperçu des travaux existants dans le domaine de la question-réponse. Notamment, il prend en compte une recherche bibliographique de la question-réponse dans plusieurs langues telles que l'anglais, le chinois, la japonaise, le français et l'arabe. Dans un premier temps, nous présentons les motivations des chercheurs pour développer de nouveaux systèmes de question-réponse pour l'arabe dont nous constatons un peu de travaux proposés. La question-réponse admet une préocupation particulière en d'autres lanues par rapport à l'arabe. Ceci est du de sa spécificité ainsi que de ses enjeux aux niveaux des techniques et outils utilisés. Nous présentons, dans un deuxième temps, une revue de la littérature de différents systèmes proposés pour plusieurs langues. Par la suite, nous envisageons d'apporter une nouvelle approche et un nouveau système dédiés à la question-réponse arabe. Nous présentons également une analyse performante des études proposées dans la littérature de la question-réponse arabe. Nous finirons ce chapitre par aperçu sur quelques avantages et inconvénients des systèmes de question-réponse.

1. Motivations pour un système de question-réponse

Avec l'expansion des médias électroniques, notamment le Web qui est devenu la principale source d'information pour tout le monde, la quantité d'information a augmenté de façon exponentielle. A l'existence de cette numérisation à large échelle, trouver une information de haute précision est un défi. En plus, les moteurs de recherche comme Google, probablement le plus avancé de tous, aident les utilisateurs à trouver les informations pertinentes en se basant sur des mots-clés, rechercher et récupérer un grand nombre de liens. Ces moteurs laissent à la charge de l'utilisateur de procéder à un tri important, ce qui requiert un effort conséquent de sa part et une perte de temps considérable, voire qui peut l'induire en erreur. De ce fait, avec les moteurs de recherche ou les systèmes de recherche d'informations actuels, donner une liste de mots clés ne permet de retourner qu'une liste de documents pertinents qui contiennent ces mots [Allam & Haggag, 2012]. Cependant, dans de nombreux cas, aucun des documents récupérés ne contient la réponse désirée. Dans plusieurs cas, les moteurs de recherche présentent des performances relativement élevées pour la recherche d'information. Néanmoins, elles s'avèrent inefficaces en cas de recherche d'information précise. En plus, les utilisateurs ne sont pas seulement intéressés à obtenir les pages pertinentes, mais ils sont souvent intéressés à obtenir une réponse précise à une question donnée [Hirschman & Gaizauskas, 2001], [Zhang & Lee, 2003]. Alors, avoir une information

précise devient une tâche ardue. En réalité, la recherche d'une information ou d'un document sur le Web est devenue une activité quotidienne. Pour pallier les problèmes des moteurs de recherche qui sont aptes de retourner plusieurs documents, des systèmes de question-réponse ont émergé comme une solution alternative à fournir des réponses pertinentes et précises à des questions des utilisateurs.

Dans ce cadre, plusieurs chercheurs, tels que [Cao et al., 2010], [Wren, 2011], [Bauer & Berlant., 2012], [Mishra & Jain, 2015] soutiennent l'idée que la question-réponse est une forme spécialisée de recherche d'informations, elle est non seulement intéressée à obtenir des pages pertinentes, mais à obtenir des réponses spécifiques à une liste de questions en langage naturel. Ces auteurs suggèrent qu'un système de question-réponse est un type particulier dédié pour la recherche d'information précise. Le résultat de ce système consiste en quelques mots contenant la réponse. En effet, des paradigmes de recherche d'informations, tels que Google ou Yahoo permettent de retourner un ensemble de pages censées contenir l'information recherchée à partir de mots clés définissant la recherche. Toutefois, l'utilisateur doit trouver les mots clés pertinents ce qui n'est pas toujours facile. Il doit également examiner les passages de texte et parfois même les pages censées contenir l'information recherchée ce qui peut être coûteux en temps. Par conséquent, ces contraintes majeures donnent lieu à l'apparition des systèmes de question-réponse, qui engendrent des solutions [Grappy, 2011]. Ou alors, lors d'une recherche d'information, l'utilisateur connaît a priori le sujet qui l'intéresse. En revanche, cet utilisateur ne connait pas sous quelle forme il peut trouver cette information, ni comment formuler sa demande d'informations.

De leur part, [Benamara, 2004] préconise qu'avec l'hétérogénéité des contenus du web, leur diversité et leur étendue, les exigences des utilisateurs croissent en parallèle. En employant les moteurs de recherche comme Yahoo, Google, etc., la réponse est un ensemble de liens vers des pages ou des portails Internet. Ces pages incluent de nombreuses redondances et qu'elles probablement ne répondent pas nécessairement à la question posée. Ainsi, la faiblesse des systèmes d'indexation fondés sur des mots-clés font que, au court terme tout au moins, il n'y a pas d'alternative technique viable. Il semble donc intéressant d'introduire le paradigme des systèmes de question-réponse qui sont conçus à trouver uniquement des portions de documents qui répondent à la question.

Sachant que l'apparition des systèmes de question-réponse a été introduite via le manque de précision dans la recherche documentaire. Dans ce cadre la recherche d'informations a évolué pour trouver des définitions, des références d'articles scientifiques, etc. Ceci se fait grâce aux systèmes de question-réponse dont la réponse doit être extraite ou recomposée à partir de documents [Perret, 2005]. D'ailleurs, trouver une information dans des grands ensembles de données hétérogènes amène de nouveaux problèmes. Néomoins, avec un système de recherche d'informations performant, une requête représentant fidèlement le besoin d'informations et le temps pour fouiller à travers une liste de documents considérés pertinents, trouver cette information sera plus facile [Belanger, 2006].

Plusieurs autres études telles que [Kor, 2005], [Stenchikova et al. 2006], [Grau et al., 2005] utilisent les moteurs de recherche pour une première sélection de documents à base des mots clés des questions. Ces moteurs sont d'une efficacité incomparable, cependant la finesse des résultats obtenus n'est pas toujours idéale. Pour faire face à ce constat, plusieurs travaux qui sont intéressés au développement de nouveaux paradigmes de recherche adaptés (e.g. les systèmes de question-réponse ont été effectués). Dans ce cadre, pour chercher une information, il faut donner un ensemble de mots-clefs. Ces mots sont des index, ils sont jugés pertinents pour trouver l'information demandée. La réponse du système est un ensemble de liens vers des pages Internet [Garcia-fernandez, 2010]. Par exemple, dans le système FRASQUES [Grau et al., 2005], la recherche et le traitement des documents sont effectués par le moteur de recherche Lucène[4]. Par conséquent, les paramètres donnés aux moteurs de recherche sont issus de l'analyse de la question : mots de la question et leurs variations. Nous présentons aussi un autre système nommé AnswerBus [Zheng, 2002] qui répond aux questions des utilisateurs en utilisant cinq moteurs de recherche et annuaires (Google, Yahoo, Wisenut, AltaVista et Yahoo News) pour récupérer des pages Web qui contiennent potentiellement des réponses. Par ailleurs, QASR [Stenchikova et al. 2006] est un système de question-réponse en domaine ouvert cherchant les réponses à partir du web. Ce système utilise Google pour extraire du web les documents contenant potentiellement la réponse précise. Pour le faire, les auteurs utilisent un annotateur de rôles sémantiques [Pradhan, et al. 2005]. Les rôles sémantiques correspondent à la relation existant entre un prédicat et un constituant syntaxique.

[4] https://lucene.apache.org/

La discipline de la question-réponse a été développée depuis les années 1960. Cette discipline favorise des systèmes pour plusieurs domaines, des sources de données différentes, des types de questions variés, des formats spécifiques des réponses, etc.; le nombre des systèmes proposés chaque année est trop elevé. Ces systèmes sont expérimentés pour plusieurs langues. En effet, l'anglais a été considérée la langue la plus étudiée dans ce domaine. Par la suite, la question-réponse a été etudiée dans d'autres langues, le chinois, le japonais, etc. Dés lors, la question-réponse a introduit de nouvelles langues, les langues européennes ont été essentiellement étudiées dans CLEF, les langues asiatiques dans NTCIR, la langue française dans EQueR, la langue arabe dans TREC et CLEF, etc. Par conséquent, pour relever les méthodologies de ces systèmes et leur capacité à satisfaire les besoins des utilisateurs pour les informations précises, une enquête systématique de tous ces des paradigmes ainsi que leurs approches devient nécessaire.

2. Aperçu de la question-réponse en d'autres langues que l'arabe

La question-réponse est un domaine de travail multidisciplinaire dont un système de question-réponse incorpore souvent des techniques et des ressources éventuelles, y compris la recherche d'informations, le TALN, l'extraction d'information, l'apprentissage automatique, etc., [Lee et al. , 2005]. Les recherches dans ce domaine d'étude ont vu le jour depuis les années 1960. Il existe des méthodologies et approches considérables qui ont été présentées pour plusieurs langues (bulgare, néerlandais, anglais, français, allemand, indonésien, italien, japonais, portugais, arabe et espagnol, etc.). La question-réponse a souvent été produite dans des compagnes et des ateliers de validation, tels que, TREC, CLEF, NTCIR, MUC, etc. Il ne se trouve pas d'approche standard, à l'heure actuelle, pour les systèmes de question-réponse. De ce fait, les architectures des systèmes peuvent être différentes. Néanmoins, nous pouvons noter une structure commune ou architecture typique pour un système de question-réponse qui se décompose généralement par trois ou quatre étapes.

Les systèmes de question-réponse peuvent être différenciés selon leurs approches, techniques et outils étudiés. Dans ce qui suit, nous présentons quelques approches caractéristiques qui ont obtenu les meilleurs résultats dans les tâches de question-réponse lors des récentes campagnes d'évaluation TREC, CLEF et EQueR, etc. Par conséquent, depuis l'inclusion de technologie de la question-réponse dans des conférences et des workshops dés TREC-8 en 1999, nous constatons une croissance dramatique des systèmes de question-réponse au niveau théorique et empirique. En outre, l'anglais était jadis considéré comme la

langue universelle, particulièrement sur la Toile. La question-réponse pour la langue arabe a vu ses premiers développements en 1990. De plus, la nécessité de considérer plusieurs langues simultanément a été aperçue dans les pays de l'union européenne. Dés lors, la question-réponse devait évoluer pour traiter de nouvelles langues. Cependant, peu d'études ont porté sur l'arabe. La pénurie des travaux pour la langue arabe peut, en quelque sorte, explique la difficulté de cette langue.

2.1 Question-réponse en anglais

Il y a une riche bibliographie de la question-réponse en anglais. Par conséquent, c'est à priorité pour cette langue que la question-réponse a connu ses premiers développements, de nombreux travaux ont été proposés pour la recherche des réponses à des questions. Généralement, l'entrée des systèmes est une question en langage naturel et la sortie est une liste des entités concernées. Ainsi, une revue de la littérature révèle que dans le traitement des langues naturelles et de recherche d'informations, l'anglais est la langue la plus étudiée en termes de ressources, corpus et systèmes [Ligozat, 2013].

La conférence TREC (Text Retrieval Evaluation Conference) introduit la piste de question-réponse en 1999. Quelques années après, spécialement en TREC 2004, le meilleur système de question-réponse permet de s'exécuter avec une précision de 77% a fin répondre à des questions factuelles [Voorhees, 2004] (e.g. How many calories are there in a Big Mac?) et atteint un score de 62.2% pour répondre à des questions liste (e.g List the names of chewing gums).

Initialement, les premiers systèmes en anglais sont apparus dans les années 60 grâce à l'introduction des approches fondées sur le dialogue Homme-Machine. De ce fait, il ya eu plusieurs enquêtes sur la technologie de la question-réponse pour cette langue. Auparavant, Simmons a examiné les premières approches pour répondre à des questions en anglais [Simmons, 1965]. En effet, l'intérêt du développement des systèmes de question-réponse a commencé depuis la tentative faite par [Green et al., 1961] à travers le système BASEBALL. De plus, les premiers systèmes mettent l'accent sur une analyse en profondeur de la question. Quelques systèmes on été proposés en domaine ouvert, d'autres travaux ont porté sur la recherche de réponses en domaine de spécialité. L'approche utilisée dans ces systèmes reposait sur la transformation d'une question posée en langage naturel en une requête afin de récupérer une réponse courte à partir de la base de données interrogée. Par ailleurs, les

systèmes existants considèrent différents types de questions allant des plus simples (e.g. les questions factuelles) aux plus complexes (e.g. les questions pourquoi, les questions d'opinion, etc.). Ces systèmes nécessitent toujours des approches plus profondes.

D'abord, les approches les plus utilisées étaient à base de surface (e.g. les techniques statistiques ou symbolique / pattern matching, etc.). L'anglais tient parti plus de ressources linguistiques par rapport aux autres langues. Ainsi, une approche qui est fortement fondée sur des ressources ne serait pas abordable à une autre langue non dotée de mêmes ressources. D'ailleurs, des investigations ont montré que le pourcentage de pages Web qui sont exprimées presque en anglais commence à inclure d'autres langues comme le chinois [Global-research, 2001].

Ainsi, de nombreux systèmes s'appliquant sur l'anglais ont recours à WordNet [Fellbaum, 1998] pour détecter le rapprochement entre deux mots. La méthode la plus simple, présentée notamment dans [Pakray et al. 2009] ou [Ofoghi, 2009], est fondée sur les relations de synonymie ou d'hyperonymie. Certaines approches n'auront pas la même efficacité selon la langue, le type de document traité, les ressources linguistiques à disposition, etc. Ainsi, PowerAnswer est un système de question-réponse proposé par [Moldovan et al., 2002] du LCC (Language Computer Corporation) avec une architecture fondée sur le raisonnement logique, il repose sur la représentation sous forme de formules logiques de la question, de la réponse ainsi que des sources de données servant à extraire la réponse. La question est analysée afin de déterminer son type, le type de la réponse attendue et les mots-clés qui la composent. Cette analyse utilise les données sémantiques de WordNet ainsi identifie les entités nommées présentes dans la question. Par ailleurs, ce système donne des résultats encourageants sur l'anglais. Cependant, les approches utilisées dans ce système sont difficilement reproductibles dans d'autre langues, car Chaucer fait appel à des ressources linguistiques de grande envergure disponibles uniquement pour l'anglais : FrameNet [Ruppenhofer, et al. 2006], Extended WordNet [Harabagiu, et al. 1999].

Une autre approche utilisée par [Stenchikova et al. 2006] s'appuie fortement sur un annotateur de rôles sémantiques pour la mise en considération le système QASR qui cherche des réponses à des questions sur Internet. Ainsi, YourQA est un système de questions-réponses en anglais [Quarteroni & Moschitti, 2010]. Ce système utilise le Web, il est en

domaine ouvert. Ce système exploite principalement des combinaisons de noyaux pour réordonner les réponses candidates retournées par l'extracteur de réponses.

Dans le laboratoire LIMSI[5], QALC (Question Answering program of the Language and Cognition group) [Ferret et al., 2000], [Ferret et al., 2001a], [Ferret et al., 2001b] et [Ferret et al,. 2002a], est considéré comme le premier système de question-réponse en anglais, développé dans le contexte de la compagne d'évaluation TREC (Text REtrieval Conference) 1999. Ce système génère une réponse à une question à partir d'un grand volume de documents. L'idée de base de ce système est de trier parmi plusieurs phrases candidates, les 10 premières qui assurent une réponse convenable et précise à la question posée. Cette idée garde son efficacité si les phrases sélectionnées possèdent de poids différents. Dans le cas contraire, la chance de trouver la réponse précise peut se trouver aussi bien dans l'une des dernières phrases que dans la première phrase.

En 2012, QALC a subi une adaptation d'un système existant pour une tâche de compréhension automatique. A cet égard, Grau et ses collègues ont proposé une approche de sélection de réponses correctes en se basant sur la reconnaissance de l'implication textuelle entre les textes et les hypothèses [Grau et al., 2012]. Tandis que, QALC a été conçu au début pour répondre à des questions factuelles dans n'importe quel domaine. Le processus d'extraction de la réponse est décrit comme suit :

- Utiliser les documents sélectionnés par les moteurs de recherche,
- Séparer les phrases à fin de comparer chaque phrase à la question,
- Enfin, localiser la réponse extraite que ce soit par la détection des entités nommées ou par application des patrons d'extraction de la réponse.

Il y a eu récemment d'autres études qui ont été menées sur cette ligne de recherche en favorisant des systèmes de question-réponse en ligne. En effet, AskHERMES[6](Ask Help clinicians to Extract and aRticulate Multimedia information for answering clinical quEstionS) permet aux médecins de poser des questions complexes et médicales aux moments des consultations des patients et d'identifier rapidement des réponses précises et exactes à ces questions [Cao et al., 2010]. Ce système aide d'une part les médecins à extraire et de formuler des informations multimédia à fin de répondre à des questions médicales. D'autre part, il

[5] http://www.limsi.fr/
[6] http://www.askhermes.org/

assure l'amélioration de la qualité des soins. AskHERMES permet également d'analyser de gros volumes de documents relatifs à des questions pour produire de textes courts comme réponses. Les auteurs ont évalué leur système par 4654 questions médicales collectées en pratique. Les résultats affichés dans [Cao et al., 2010] montrent que Cao et ces collègues ont réalisé les deux scores suivants 76.0% et 58.0% conçus successivement pour la tâche de classification des thèmes généraux et la tâche d'extraction des termes. Ce système répond aux exigences des médecins en consultation tout en répondant à leurs questions médicales ad hoc dans une période de temps. Il assure l'insertion des données de la littérature ou d'autres sources d'informations pour répondre à ces questions.

Les succès rencontrés par les systèmes de question-réponse en domaine ouvert, au nombre desquels il faut compter la performance récente du système Watson d'IBM [Ferrucci et al., 2010], ne doivent néanmoins pas cacher leurs limites dès lors qu'ils sont appliqués à des domaines plus spécialisés. Qakis [7] avait l'abréviation de (Question Answering wiKiframework-based System), un système de question-réponse en domaine ouvert. Il génère des requêtes SPARQL à partir des questions, les soumet à DBpedia[8] et compare la question à la base de modèle pour identifier la relation entre DBpedia et identifier la réponse. En outre, il fournit une intégration des chapitres multilingues à une requête DBpedia avec des requêtes en langage naturel. QAKiS permet aux utilisateurs de soumettre une requête à un magasin triple RDF en anglais et obtenir la réponse dans la même langue. Son architecture est composée de quatre éléments (génération de requête, reconnaissance d'entités nommées, appariement de formes et package SPARQL) [Cabrio et al., 2012]. À l'instar de ce que fait Qakis pour les requêtes SPARQL, un autre système de question-réponse, appelé SELNI, a été conçu. Ce dernier utilise l'algorithme SVM pour former des requêtes SPARQL ainsi des questions factuelles pour interroger le serveur DBpedia afin d'en extraire la réponse exacte. Celui-ci est basé sur l'ontologie DBpedia Infobox [Tahri & Tibermacine, 2013]. D'ailleurs, SELNI a atteint une précision de 86% sur la base de l'ensemble de test de TREC 10.

Un autre système qui mérite une attention particulière dans la question-réponse en anglais, c'est celui proposé par [Bhaskar et al., 2012]. Ce dernier participe à la tâche principale de QA4MRE@CLEF 2011 et QA4MRE@CLEF 2012. Il utilise plusieurs mesures d'implication textuelle (comparaison des entités nommées, n-grammes et skip n-grammes

7 http://qakis.org/qakis/

8 http://dbpedia.org

communs) et compare le type de réponse avec celui attendu. Ceci est le système qui a obtenu les meilleurs résultats à la tâche générale de QA4MRE 2011 et 2012 @ CLEF. Dan ce cadre, les auteurs combinent la question et chaque option de réponse pour former l'hypothèse (H), les mots vides ont été supprimés à partir de chaque hypothèse. Chaque phrase permet de définir le texte T et la paire (T, H) alloue un score classement sur la base reconnaissance d'implication textuelle. Ainsi, chaque phrase est attribuée un score d'inférence par rapport à chaque modèle de réponse. La réponse choisie est celle qui reçoit le score le plus élevé dans la liste des options de réponses. Ce système a une précision de 0,53 et c @ 1 de 0,65.

2.2 Question-réponse en chinois et japonais

Le chinois est la deuxième langue la plus populaire dans la technologie de la question-réponse. Elle a été introduite pour la première fois en 2005 à NTCIR [Lee et al., 2005]. Plusieurs études ont été attaquées par la tâche de génération des réponses précises à des questions. En fait, Marsha, un système de question-réponse, repose sur les mêmes techniques utilisées dans les systèmes anglais développés par TREC [Li & Croft, 2001]. Marsha est porté sur une méthode basée sur les questions du TREC. Il a la même performance que certains systèmes de question-réponse en anglais à la piste TREC-8. D'ailleurs, Li et Croft ont utilisé 51 questions. Ils ont sélectionnés 26 questions à partir de 240 recueillies par des étudiants chinois. Le reste de ces questions sont spécifiées à reformuler une question ou demander autres légèrement différentes. Marsha contient des techniques spécifiques traitant avec les caractéristiques chinoises (segmentation de mots, traitements ordinaux). Il est composé de trois modules, tels qu'un traitement des questions, un moteur de recherche et un module d'extraction de la réponse.

En outre, autres chercheurs ont traité des questions anglaises et chinoises. Ils étudient une conversion anglais-chinois et chinois-anglais via des systèmes de traduction automatique. Ils participent également à la tâche CLQA (Cross Language Question Answering) [Kwok et al., 2005] de NTCIR. Dans un autre travail de [Lee et al.,2005], les auteurs exposent un système ASQA (Academia Sinica de question-Answering). Ce dernier s'occupe de fournir des réponses chinoises à partir de questions factuelles. Ce système est fondé sur une architecture hybride évaluée par six types de questions factuelles tels que des noms de personnes, des noms des lieux, des noms d'organisations, des objets, des dates et des nombres. ASQA combine des approches qui sont fondées sur l'apprentissage automatique et la connaissance. Il

a atteint des valeurs de précision Top1 37,5% et 44,5% respectivement pour des réponses correctes et non supportées.

Apart le chinois et l'anglais, dans NTCIR-6, Mitamura et ses collègues ont participé à quatre sous-tâches de CLQA (J-J, E-J, C-C et C-E[9]) et introduisent la langue japonaise [Mitamura et al., 2007]. Par exemple, [Nyberg et al., 2002] ont proposé un système de question-réponse JAVELIN (Justification based Answer Valuation through Language INterpretation) fondé sur une interaction avec l'utilisateur dont l'intérêt général est d'élucider la question et de déterminer une stratégie de recherche adaptée pour trouver la réponse. Avec le système JAVELIN les réponses à des questions en anglais sont extraites à partir de documents japonais et chinois. JAVELIN III était une extension de leur système précédent JAVELIN II qui a été initialement conçu pour question-réponse monolingue anglais [Nyberg et al., 2005]. Ainsi, JAVELIN III est spécifié par un système modulaire, extensible et doté par une architecture indépendante de la langue. Une meilleure exécution a obtenu une précision de 13% pour E-J et une précision de 19% pour les C-E.

Récemment, NTCIR-9 introduit pour la première fois la nouvelle tâche RITE @ NTCIR (Reconnaissance d'inférences textuelles). RITE est une tâche générique qui gère une compréhension majeure du texte dans divers domaines de recherche, comme la recherche d'information, question-réponse [Harabagiu & Hickl, 2006], récapitulation de textes, analyse de l'opinion, etc. L'idée est d'assurer des progrès dans la recherche d'inférence textuelle [Shima et al., 2011]. Selon ces auteurs, la sous-tâche RITE4QA est inspirée par une série de tâches de validation de la réponse au CLEF [Peñas et al., 2007].

2.3 Question-réponse en français

Un système de question-réponse vise à analyser des questions et extraire des réponses pertinentes et courtes localisées dans de petits fragments de texte au lieu de donner des documents ou des passages de textes comme le donnent les systèmes de recherche d'informations [Mervin, 2013]. En 2004, il aura lieu la première évaluation des systèmes de question-réponse en français, EVALDA-EQUER, fondée par le Ministère de la Recherche, dans le cadre de l'Action Technolangue[10]. La campagne proposait deux tâches, une première dans le domaine général et une deuxième tâche plus spécialisée. De ce fait, beaucoup de

[9] J-J, E-J, C-C et C-E : japonais-japonais, english-japonais,chinois-chinois et chinois-english
[10] http://www.technolangue.net/article20.html

systèmes de question-réponse élaborent des résultats favorables s'il s'agit de chercher des documents mais ils possèdent plus de contraintes de trouver une réponse correcte en première position.

Par exemple, dans [Grau et al., 2006a] les auteurs ont développé un système de question-réponse, FRASQUES. Ceci est, en réalité, une version adaptée au français du QALC qui est dédié pour l'anglais. En outre, les ressources sur lesquelles porte l'étape d'analyse des questions (analyseur syntaxique, étiqueteur morphosyntaxique, etc.) ont été rectifiées. Les sorties de ces outils ont été projetées sur des formats communs en anglais et en français, afin que le module d'analyse des questions puisse être le même dans les deux langues. Le système REVISE a été conçu sur la base du modèle du système de question-réponse réalisé par [El Ayari et al., 2009] et nommé FRASQUES.

Un autre système de question-réponse appelé WEBCOOP (COOPérativité pour le WEB), proposé par [Benamara, 2004], il génère des réponses coopératives. L'idée est de proposer à l'utilisateur des informations additionnelles (explications, justifications, etc.). Ce système de question-réponse trouve une réponse même si la question posée comporte des fausses présuppositions ou des malentendus. WEBCOOP repose sur l'intégration de procédures de raisonnement couplées à des modes de représentation de connaissances.

Dans le cadre des systèmes de question-réponse en domaine restreint, la recherche documentaire se fait sur un ensemble généralement limité de documents alors que pour les systèmes en domaine ouvert, la recherche d'informations s'effectue sur une grande collection de textes couvrant presque tous les domaines, à savoir, les sources de données existantes sur le Web. Par exemple, Esculape [Embarek & Ferret, 2010], est un système de question-réponse en français conçu pour les médecins généralistes et constitué à partir d'Œdipe. Ce dernier est un système de question-réponse en français et en domaine ouvert. Esculape ajoute à Œdipe la capacité d'exploiter la structure d'un modèle du domaine, le domaine médical dans le cas présent. En fait, Œdipe [Besançon et al., 2007] est un système de question-réponse français reprenant l'architecture classique d'un système de question-réponse en domaine ouvert mais il est conçu comme étant une extension minimaliste d'un moteur de recherche.

Avec l'usage du Web comme une source de connaissances, les systèmes de question-réponse souffrent de la redondance informationnelle [Demner-Fushman & Lin, 2007]. Cependant, la fiabilité de ces informations est mise en cause. En effet, le développement du

Web et l'amélioration immense des outils de traitement automatique du langage naturel ont largement contribué à la possibilité de réaliser des systèmes de question-réponse ayant pour objectif de répondre à tout type de questions. Dans ce contexte, plusieurs systèmes de question-réponse ont vu le jour. En effet, QRISTAL (Questions-Réponses Intégrant un Système de Traitement Automatique des Langues) [Laurent, et al. 2010] est un système de question-réponse multilingue (français, anglais, portugais, italien et polonais), développé par Synapse Développement, pour extraire des réponses dans une base documentaire locale ou à partir du Web. Ce système a obtenu les meilleurs résultats sur le français, il a effectué 68 % de bonnes détections sur les données de la campagne d'évaluation CLEF 2006 [Magnini et al. 2006]. De même, Citron [Falco, 2014] est un système de question-réponse en français capable d'extraire des réponses à des questions à réponses multiples en domaine ouvert à partir de documents provenant du Web. La particularité de Citron est qu'il exploite les structures (énumérations, tableaux) propres à ces documents pour mieux extraire les réponses et éventuellement les agréger pour faire apparaître des critères variantes comme la date ou le lieu.

Parfois, les systèmes de question-réponse sont inter-lingue, cela signifie que la langue des documents est différente de la langue des questions, comme c'était le cas pour le système MUSCLEF qui a également été développé par [Grau et al., 2006]. Depuis 2005, ce système a participé aux campagnes d'évaluation QA@CLEF en 2005 et 2006. MUSCLEF prend en entrée des questions en français, et recherche leurs réponses dans des documents en anglais. Pour construire ce système, les auteurs ajoutent certains modules à leur système QALC en mettant en considération deux stratégies possibles pour faire face l'inter-lingue, la traduction de la question ou la traduction terme à terme. Pour chacune de ces stratégies, les auteurs utilisent d'autres sources externes pour la tâche de la traduction.

La formulation des réponses en langue naturelle a reçu plus d'attention, permettant ainsi la création de systèmes interactifs [Mendes & Véronique, 2004]. Dans un dialogue, les interlocuteurs à la recherche d'informations ne souhaitent pas forcément connaître la source, ni les justifications des réponses, mais plutôt obtenir une réponse en langage naturel favorisant ainsi une interaction entre eux. Un des systèmes les plus connus qui suit ce type d'approche est le système RITEL [Galibert, 2009] ; [Bernard et al., 2009]. Ce système répond à des questions en domaine ouvert, il se base sur une analyse complète et multi-niveaux de questions et de documents. Il est apparu au LIMSI en 2004 [Galibert, et al., 2005]. Ainsi, il est

désigné initialement à un système de dialogue home-machine [Toney et al., 2008]. En outre, RITEL sert à offrir à l'utilisateur la possibilité d'interroger avec un système de recherche d'informations générale via un navigateur web. Cette interrogation est effectuée en en incluant des capacités conversationnelles et orales. Par conséquent, avec RITEL les questions et les documents sont analysés puis indexés.

Dans la littérature de question-réponse, la plupart des systèmes ont été conçus pour répondre à des questions factuelles qui attendent généralement comme réponse une entité nommée, peu des systèmes se sont intéressés pour répondre à des questions complexes. Le terme de question complexe est défini comme regroupant les questions non factuelles, même si ce critère varie. Parmi les catégories de questions complexes, nous citons celles de liste, questions de raison, des questions comment et pourquoi, etc. Malgré la difficulté de ces questions, nous présentons également des expériences préliminaires avec de bons résultats. Par exemple, le travail proposé par FIDJI [Moriceau et al., 2010] se situe dans le cadre de développement du système de questions-réponses FIDJI développé pour le français et l'anglais et qui traite des questions factuelles et complexes (pourquoi et comment), en combinant des informations d'ordre syntaxique et des techniques "classiques" du domaine, telles que la reconnaissance des entités nommées et la pondération des termes de la question.

Certains autres systèmes portent sur la tâche de validation. Selon une définition formulée par [Grappy, 2011], une réponse est valide si elle est compatible avec la question et que les informations demandées par la question se trouvent dans le passage justificatif. Comme exemple de systèmes qui appliquent une méthode de validation de réponses, QAVAL (Question Answering by VALidation) qui est développé par [Grappy et al., 2011]. Il utilise le moteur de recherche Lucene[11] pour sélectionner des passages courts au lieu des documents. Ces passages sont analysés par un analyseur terminologique un peu profond. En outre, QAVAL se fonde sur FASTR [Jacquemin, 1999] à fin d'extraire et de reconnaître les termes et ses variantes. Les meilleurs passages sont choisis en fonction de la présence des termes de la question. Ainsi, les réponses candidates sont sélectionnées de ces passages déjà analysés. Enfin, une machine d'apprentissage permet d'appliquer plusieurs critères pour la tâche de classification de ces réponses. L'évaluation du QAVAL se base sur un corpus web construit

[11] http://lucene.apache.org/core/index.html

de la société Exalead[12]. Egalement, QAVAL applique une méthode de validation de réponses. Cette dernière consiste à ordonnancer ces réponses extraites à partir de 300 passages de textes.

3. La question-réponse en arabe

Dans les sections précédentes, différents systèmes et approches autorisent d'obtenir des réponses précises à des questions en langues latines ont été présentés. En effet, depuis son émergence, la question-réponse a réalisé une performance globale dans ces différentes langues, particulièrement pour l'anglais et quelques autres langues latines qui sont beaucoup plus bénéficiées de l'avancement dans le domaine de la question-réponse. Toutefois, les études qui ont été proposées pour l'arabe ne sont pas aptes de suivre le même rythme d'évolution en raison des défis spécifiques de cette langue. D'ailleurs, la plupart de ces études avaient porté sur des techniques de TALN pour extraire la réponse exacte. Etant donné que, d'après une revue bibliographique, nous remarquons que la quasi-totalité de ces systèmes de question-réponse en arabe traitent des approches morphosyntaxiques et que peu d'entre elles fournissent des approches fondées sur la sémantique, l'inférence et la logique. Mais, malgré ce manque, il existe dans la littérature quelques tentatives qui ont obtenu des résultats acceptables dans ce domaine de recherche. A notre connaissance, la technologie de question-réponse arabe a été étudiée, entre autres, depuis les années 1990 (figure 2.4). De ce fait, il est nécessaire d'examiner ces revues afin de noter les différentes expérimentations obtenues par chaque travail.

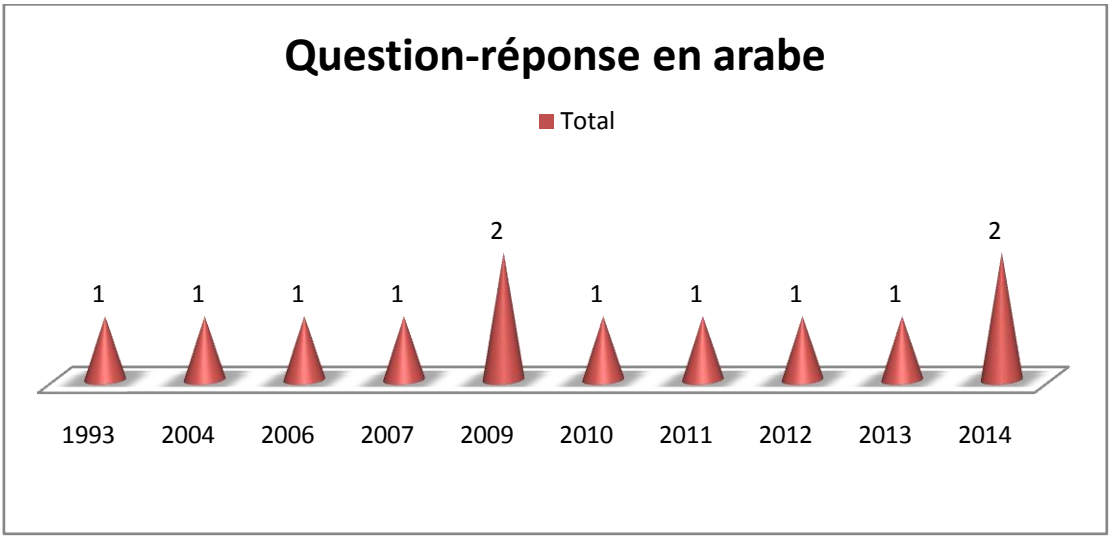


Figure 2.4: Evolution de la question-réponse en arabe depuis son apparition

[12] http://www.3ds.com/products-services/exalead

La nécessité des systèmes de question-réponse accroît dans le cadre de la langue arabe. Ceci est dû à la pénurie de ces systèmes, qui peut être attribuée aux grands défis qu'ils présentent à la communauté des chercheurs, y compris la spécificité de la langue arabe, tels que les voyelles courtes, absence de lettres majuscules, morphologie complexe, etc., [Kurdi et al., 2014]. C'est pratiquement à partir de 2004, que quelques tentatives exploratoires ont été réalisées autour des systèmes de question-réponse arabes. Ces études seront bien présentées dans la suite de cette section.

3.1 Les défis de la langue arabe

L'importance de la technologie de la question-réponse est très évidente et pertinente en matière de recherche d'informations. Cette technologie a été établie pour plusieurs langues latines telles que l'anglais, le français [Akour et al., 2011], [Bekhti & Al-Harbi, 2013], etc. Néanmoins, les systèmes de question-réponse accomplis en arabe sont encore peu nombreux et immatures en raison des aspects uniques de la langue arabe. Ceci est principalement dû à deux raisons: le manque d'accessibilité aux ressources et aux outils linguistiques (e.g. les corpus et les outils de TALN arabes) et la nature très complexe de la langue elle-même (e.g. l'arabe est flexionnelle et non-concaténative). Bien que, l'arabe est dans les dix premières langues dans l'Internet, il manque de nombreux outils et ressources. L'arabe n'a pas les lettres majuscules par rapport aux langues latines, comme dans le cas de l'anglais. Ce problème rend si dur le traitement du langage naturel, comme la reconnaissance des entités nommées. En outre, l'arabe est l'une des langues les moins considérées par les chercheurs dans le champ de la question-réponse [Abouenour et al., 2012].

De leur côté, [Abdelnasser et al., 2014] illustrent quelques difficultés de la langue arabe. En effet, ce langage est très flexionnel et dérivationnel ce qui rend son analyse morphologique une tâche très complexe. D'abord, dérivationnel où tous les mots arabes ont trois ou quatre caractères racines. Ensuite, flexionnel où chaque mot se compose d'une racine et zéro ou plusieurs affixes (préfixe, infixe, suffixe). L'arabe se caractérise par des marques diacritiques (voyelles courtes), le même mot avec différents diacritiques peut exprimer des significations différentes. Les diacritiques qui provoquent l'ambiguïté sont généralement omises. Ainsi, l'absence des lettres majuscules en arabe est un obstacle contre la reconnaissance des entités nommées.

Par conséquent, nous illustrons dans divers travaux [Abufardeh & Magel, 2008], [Al-daimi & Abdel-amir, 1994], [Habash & Rambow, 2005], [Akour et al., 2011] qu'il existe plusieurs facteurs motivants pour choisir la langue arabe. Par la suite, nous citons quelques exemples de ces facteurs :

- La langue arabe est la sixième langue la plus parlée dans le monde.
- La langue arabe a approximativement 280 millions de parlants natifs et environ 250 millions de parlants non-natifs.
- La langue arabe est considérée également l'une des six langues officielles des Nations Unies (anglais, arabe, chinois, anglais, français, russe et espagnol).
- Croissance des données textuelles arabes sur le web et l'évolution des demandes de logiciels arabe de haute qualité.
- La langue arabe est hautement flexionnelle et dérivationnelle, ce qui rend l'analyse morphologique une tâche très complexe.
- Pas de capitalisation en arabe. Cela complique le processus d'identification des noms propres, acronymes et abréviations.
- La direction d'écriture est principalement mélangée de droite à gauche et de gauche à droite. En outre, certains caractères changent leurs formes en fonction de leur emplacement dans un mot.

3.2 Principaux systèmes proposés

De nos jours, la plupart des systèmes de question-réponse traitent des questions en langues latines (l'anglais, le français, le chinois, le japonais, etc.). Le besoin de développer des systèmes de question-réponse dédiés à l'arabe devient de plus en plus inévitable ces dernières années à cause des difficultés liées à la langue elle-même aussi par le manque d'outils disponibles pour aider les chercheurs. Par conséquent, les approches proposées signalent qu'il n'y a pas vraiment une méthode standard à adopter lorsqu'il s'agit d'extraire les bonnes réponses à une question. Tout va dépendre du type de problèmes que nous voulons le traiter et du cadre de recherche dans lequel nous nous voulons inscrire. Dans cette section, nous avons présenté quelques systèmes qui sont accomplis dans cette langue depuis leur naissance avec AQAS de [Mohamed et al., 1993] jusqu'à Al-Bayan [Abdelnasser et al., 2014] qui est récemment présenté.

La question-réponse arabe a vu le jour avec le système AQAS (question-answering system), présenté par [Mohammed et al., 1993]. Ceci est basé sur la connaissance, il extrait

des réponses uniquement à partir de données structurées. AQAS accepte une phrase arabe (des états déclaratifs ou une question) et génère la sortie appropriée à l'usager dans le domaine de rayonnement. Le système proposé est considéré comme étant une mise en œuvre en arabe pour le traitement du langage naturel, il est dédié pour le domaine de radiation.

L'étude des questions factuelles était le sujet de plusieurs systèmes de question-réponse, surtout en arabe. De ce fait, de nombreux systèmes tentent de trouver des entités nommées pour répondre aux questions. L'idée est que les questions factuelles se répartissent en plusieurs types distinctifs, tels que « personne », «emplacement», «date», «organisation», etc. La tâche d'identifier ces types est appelée la reconnaissance de l'entité nommée, elle est généralement effectuée par un outil ou dispositif de reconnaissance d'entités nommées. Dans ce cadre, certains systèmes en arabe adoptent une stratégie de recherche appropriée en s'appuyant sur la reconnaissance des entités nommées. Par exemple, AQUASYS (Arabic Question-Answering System) de [Bekhti & Al-Harbi, 2013] répond aux questions qui commencent par les pronoms interrogatifs (من qui, ما quoi, أين où, متى quand, كم العددية combien, كم الكمية combien). Effectivement, AQUASYS a été la base d'un autre travail connexe proposé par [Kurdi et al., 2014] pour concevoir et réaliser le système de question-réponse JAWEB (web-based Arabic question answering application system). La particularité de ce système par rapport à AQUASYS est qu'il fournit une interface utilisateur comme une extension. Ce système est axé sur le web pour répondre à des questions commençant par " متى, أين, من, ما, كم الكمية, كم العددية". Un exemple de ces questions est bien illustré dans le tableau 2.6.

Tableau 2.6: Exemple de questions répondues par JAWEB

N°	Type	Question en arabe	Question en anglais
1	WHO	من هو محمد طنجة ؟	Who is Muhammad Tangier?
2	WHEN	متى توحدت المملكة العربية السعودية ؟	When was Kingdom of Saudi Arabia united?
3	WHAT	ماهي الأهرامات المصرية ؟	What are Egyptian pyramids?
4	WHERE	أين تقع المملكة العربية السعودية ؟	Where is the Kingdom of Saudi Arabia located?
5	HOW MUCH	كم تبلغ درجة حرارة القشرة الأرضية ؟	How much is the temperature of the Earth's crust?
6	HOW MANY	كم عدد سكان الرياض ؟	How many residents are there in Riyadh?

Nous pouvons évoquer à ce sujet, un autre outil, qui est si important et qui est utilisé par [Brini et al., 2009] pour développer leur système de question-réponse QASAL (Question–Answering System for Arabic Language), c'est la plateforme NOOJ. Cette dernière est employée afin de trouver des réponses aux questions factuelles à partir d'un ensemble de livres d'éducation. Ainsi, les expérimentions réalisées par Brini et ses collaborateurs ont signalé que pour un ensemble de données de test de 50 questions le système a atteint 67,65% comme précision, 91% comme rappel et 72,85% comme F-mesure.

En réalité, l'un des facteurs clés de la réussite dans le domaine de la question-réponse est l'organisation annuelle de campagnes d'évaluation. Néanmoins, la langue arabe est approximativement absente dans la majorité de ces pistes. C'est pourquoi, les systèmes de question-réponse en langue arabe présentent de nombreux inconvénients en termes de leur processus d'évaluation. Sauf le système ArabiQA (ArabiQA: Arabic QA system) de [Benajiba et al., 2007], où leur évaluation a respecté le même pourcentage pour chaque type des entités nommées comme dans CLEF 2006. En fait, ArabiQA est doté d'une architecture générique composée de trois modules: un système de récupération de passages (JIRS), un système de reconnaissance des entités nommées arabes et un module d'extraction de la réponse.

Non seulement le type des questions factuelles à été traité dans la question-réponse arabe. A notre connaissance, le deuxième système de question-réponse présenté en arabe, est le système QARAB (Arabic Question Answering System) proposé par [Hammo et al., 2004]. Ce système est basé sur un ensemble de règles pour chaque type de question à l'exception des deux types: "ماذا, كيف" (comment et pourquoi). QARAB favorise des réponses courtes à des questions. Il cherche les réponses dans des documents non structurés extraits du journal Al-Raya. Il aborde la question comme un "sac de mots".

Mais encore, les questions de définitions sont prises en compte. Considérant qu'une question de définition est une question qui demande des informations importantes à propos de quelqu'un ou de quelque chose. A notre connaissance, en arabe le premier système qui traite ce type de question est DefArabicQA (Arabic Definition Question Answering System), proposé par [Trigui et al., 2010]. Ce système cherche les définitions candidates en utilisant un ensemble de patrons lexicaux et les catégorise en exploitant des règles heuristiques. De surcroît, DefArabicQA classe les définitions en utilisant une approche statistique. Nous pouvons mentionner que des expérimentations préliminaires portant 50 questions sur les organisations ont été menées par ces auteurs en utilisant Google et Wikipedia (tableau 2.7).

Tableau 2.7: Définitions selon le processus de correspondance et les ressources du Web

Définitions candidates	Wikipedia	Google	Totale
Technique de correspondance difficile : HM (Hard Matching)	990 (84%)	680 (92%)	1670 (87%)
Technique de correspondance tolèrent TM (Tolerent Matching)	184 (16%)	60 (8%)	244 (13%)

Les deux (HM+TM)	1174 (61%)	740 (39%)	1914 (100%)

La succession annuelle des compagnes d'évaluation de la question-réponse telles que TREC et CLEF a autorisé l'amélioration du développement des tâches plus avancées parmi lesquelles la tâche de la question-réponse pour la compréhension en lecture (QA4MRE). Cette tâche a été introduite pour la première fois en 2011 pour l'anglais dans la CLEF. L'objectif de cette tâche est de se focaliser sur la compréhension en lecture dans la question-réponse. Néanmoins, l'édition de QA4MRE @ CLEF a été étudiée pour la première fois pour l'arabe en 2012 avec certains systèmes comme IDRAAQ (Information and Data Reasoning for Answering Arabic Questions) de [Abouenour et al., 2012]. Ce système est mis en œuvre via une approche à trois niveaux afin d'améliorer la recherche des passages. Egalement, IDRAAQ couvre deux tâches très importantes : la reconnaissance d'implication textuelle et la validation de la réponse. À noter que, IDRAAQ atteint une précision de 0, 13 et c @ 1 est égal à 0, 21 sans l'utilisation des collections de base de données de CLEF. Ainsi, il repose également sur la densité du modèle de distance N-gramme, l'expansion sémantique et le WordNet arabe.

De même, un autre travail a également porté sur la tâche de QA4MRE@CLEF, ALQASIM dont son abréviation est (Question Answer Selection and Validation system). Ce système est développé par [Ezzeldin et al., 2013], il se focalise sur la sélection et la validation de la réponse et cherche des réponses à choix multiples. ALQASIM a réalisé une performance de 0,31 précision et 0,36 c @ 1 sans utiliser la collection de base de données de CLEF. De surcroît, Ezzeldin et ses associés ont comparé leur système aux trois autres proposés en 2012, un système pour l'anglais et deux autres pour l'arabe. La performance de ces systèmes via des mesures telles que c @ 1 et exactitude est présentée dans le tableau 2.8.

Tableau 2.8: Comparaison d'ALQASIM avec autres systèmes proposés en 2012

	Exactitude	**c@1**
IDRAAQ [Abouenour et al., 2012]	0.13	0.21
DefArabiQA [Trigui et al., 2012]	0.19	0.19
Système de [Bhaskar et al., 2012]	0.53	0.65
ALQASIM [Ezzeldin et al., 2013]	**0.31**	**0.36**

Dans la question-réponse arabe, les approches fondées sur l'inférence et la logique sont dans leurs premières étapes par rapport aux autres langues comme l'anglais. Au meilleur de notre connaissance, il existe peu de systèmes qui adoptent ce type d'approches. Par

exemple, [N Bdour & Gharaibeh, 2013] ont proposé un système de question-réponse basé sur la récupération de paragraphes. Il vise à récupérer les paragraphes (de longueur variable) qui contiennent des réponses à la question. Ces auteurs ont utilisé un corpus de 20 documents, et une collection de 100 questions de type oui/non. Celles-ci sont transformées en des représentations logiques. Néanmoins, nous ne trouvons pas d'informations pour la continuité de leur proposition. D'ailleurs, le manque ou l'absence de ce genre d'approches en arabe favorise la pertinence et la faisabilité de l'exploration de nouvelles approches et de nouveaux systèmes que les adoptent. C'est dans ce cadre que nous allons apporter une pierre à la proposition d'une nouvelle approche sémantique et logique pour améliorer la question-réponse arabe. Cette approche se diffère à la majorité des recherches proposées qui sont concentrées sur des aspects morphologiques et syntaxiques.

Une vision moderne de la question-réponse arabe s'intéresse à une compréhension sémantique de documents pour répondre à des questions en langue naturelle. Nous pouvons notamment mentionner le travail de [Abdelnasser et al., 2014] qui a introduit Al-Bayan dont son abréviation est (An Arabic Question Answering System for the Holy Quran), un nouveau système de question-réponse arabe qui est spécialisé pour le Saint Coran. Ce système fournit une compréhension sémantique du Coran pour répondre aux questions des utilisateurs en utilisant les ressources coraniques fiables. Il récupère les versets les plus pertinents et extrait le passage qui contient la réponse du Saint Coran et des livres d'interprétation (Tafsir), Al-Bayan atteint une précision de 85%.

3.3 Principales approches adoptées

À travers une analyse détaillée des travaux existants sur les systèmes de question-réponse arabes (section précédente), nous avons constaté que pendant plus d'une décennie, la question-réponse arabe a eu une importance par les chercheurs en mettant en considération de nouvelles approches et de nouveaux systèmes. Les approches adoptées montrent qu'il n'y a pas vraiment une méthode standard à adopter lors de la recherche des réponses exactes à une question. Tout va dépendre du type de problèmes à manipuler aussi du contexte de travail à prendre en considération. En se référant aux travaux antérieurs, nous pouvons classifier les tentatives de principales d'approches en trois parties : l'approche morphosyntaxique, l'approche hybride et l'approche sémantique. Ces approches se fondent sur une représentation de questions et de documents qui peut être très variée selon le système et les choix d'outils utilisés. Dans nos travaux de thèse, nous allons poursuivre la catégorie des approches hybrides

et nous proposons une nouvelle approche sémantique et logique pour mettre en œuvre un système de question-réponse permettant de répondre aux questions en langue arabe.

(a) Les approches morphosyntaxiques

Dans ce type d'approches, la génération d'une réponse repose extensivement sur des techniques du Traitement Automatique des Langues et de Recherche d'Information et n'implique pas forcément une analyse sémantique de la question ou de documents récupérés [Ben-abacha, 2012]. Il faut retenir que, dans l'existant des approches proposées, pratiquement toutes les recherches se concentrent sur les approches morphosyntaxiques. La recherche de réponse passe généralement par une étape d'indexation qui consiste à identifier des entités nommées ou des expressions présentant le mieux le contenu d'un document. En effet, des traitements linguistiques comme une analyse morphologique et/ou syntaxique des documents peuvent aussi enrichir ce type d'approches. D'abord, nous présentons un premier travail qui introduit une approche fondée sur le dialogue Homme-Machine [Mohamed et al., 1993]. Nous illustrons également une autre famille des travaux qui se fondent sur des approches linguistiques en utilisant la reconnaissance des entités nommées [Hammo et al., 2004], [Benajiba et al., 2007], [Bekhti & Al-Harbi, 2013], [Kurdi et al., 2014] et sur la plateforme Nooj [Brini et al., 2009]. [Trigui et al., 2010] ont proposé une approche à base de patrons lexicaux pour identifier les définitions et sur des règles heuristiques pour les filtrer. Outre la reconnaissance des entités nommées, la compréhension automatique de textes a fait l'objet de quelques travaux [Ezzeldin et al., 2013]. En effet, la compréhension automatique de textes sert à répondre à des questions en se basant sur un seul document ou un texte dont les bonnes réponses nécessitent un certain type d'inférence et un examen de bases de connaissances acquises antérieurement [Banerjee et al., 2013].

(b) Les approches sémantiques

Comme toute autre langue, l'arabe a besoin de la sémantique pour interpréter le sens de la question et des documents afin d'extraire la réponse précise et exacte. En effet, la langue est considérée comme un défi important en termes de réponses aux questions dont chaque langue bénéfice de sa propre morphologie et sémantique. En effet, une approche sémantique procède spécifiquement à une analyse sémantique de la question et des documents et produit une représentation formelle de leur signification [Ben-abacha, 2012]. Très peu de travaux ont été réalisés dans la question-réponse arabe en utilisant des approches sémantiques pour servir

l'analyse sémantique des questions ou des documents. Par exemple les travaux de [Abouenour et al., 2012] ont utilisé la distance densité du modèle N-gramme et l'expansion sémantique en utilisant WordNet arabe et même contribué à la richesse de cette ontologie. L'approche présentée dans [Abdelnasser et al., 2014] consiste à interpréter la sémantique du Coran pour répondre aux questions des utilisateurs en utilisant le Coran et ses livres d'interprétation (tafsir).

(c) Les approches hybrides

La logique est un niveau entre la syntaxe et la sémantique [Moldoan & Rus, 2001], il est le niveau le plus important et difficile en traitement du langage naturel. En se référent à cette définition, nous pouvons dire que les approches à base de la logique et l'inférence sont des approches hybrides entre celles morphosyntaxiques et sémantiques. Bien que plusieurs approches logiques aient été proposées pour d'autres langues, il ya peu de systèmes pareils pour la langue arabe. A notre connaissance, il ya une seule tentative de travaux présentée par [N Bdour & Gharaibeh, 2013]. Cette approche s'appuie sur une représentation logique des questions pour récupérer le paragraphe qui pourrait contenir des réponses pertinentes à ces questions. L'approche présentée dans nos travaux de thèse vise à combler cette lacune. En réalité, ce type d'approches a été montré pour donner de bons résultats pour l'anglais. Nous appliquons donc la même méthode pour la langue arabe. Avant de remettre en question la sélection de la réponse, nous utilisons des outils de TALN pour l'analyse morphologique et la reconnaissance des entités nommées pour analyser la question et le passage de texte récupéré à partir du Web. En effet, le recours à des approches à base de la logique et d'inférence a débuté depuis longtemps, notamment pour l'anglais. Comme exemples d'approches qui nous ont inspiré celles proposées par [Harabagiu et al., 2000], [Mollá et al., 2000], [Rinaldi et al., 2004], etc. Notre approche implique une étape de transformation logique. A ce niveau, cette étape est réalisée pour convertir les informations en langage naturel en un ensemble de prédicats logiques [Moldovan et Rus, 2001].

4. Analyse performante des systèmes de question-réponse arabes

La langue arabe possède une morphologie très complexe (caractéristiques flexionnelles et dérivationnels), les textes arabes souffrent de la rareté des voyelles à l'intérieur et de l'absence de capitalisation. Ainsi, l'arabe est non seulement importante pour le peuple arabe, mais aussi pour tous les musulmans du monde entier, car c'est la langue de

Coran [Mohamed et al., 1993]. Cependant, les recherches concernant la question-réponse sont relativement plus avancées pour l'anglais et pour d'autres langues latines par rapport à l'arabe. La section 3.2 présente quelques systèmes de question-réponse en arabe. Cette section effectue une analyse performante de ces systèmes. En réalité, cette analyse fournit une étude comparative de ces systèmes en s'appuyant sur un ensemble de critères. Une comparaison concernant les différentes tâches couvertes par ces systèmes, comme l'analyse des questions, la recherche de passage, et l'extraction des réponses, est explorée dans le (tableau 2.9). La plupart de ces systèmes suit ces trois sous-tâches. Cependant, ils peuvent être différents dans la façon dont ils mettent en œuvre tous ces sous-tâches. Comme c'est indiqué dans ce tableau, presque la majorité de ces travaux est généralement basée sur la classification des questions.

Tableau 2.9: Tâches couvertes par les systèmes de question-réponse

	Traitement de la question			Traitement des documents				Traitement de la réponse		
Tâches / Système de question-réponse	Segmentation de la question	Classification de la question	Formulation de la question	Recherche de phrases	Recherche de passages courts	Recherche de passages	Recherche de paragraphes	Extraction de la réponse	Sélection de la réponse	Validation de la réponse
AQAS [mohamed et al.,1993]				✓					✓	
QARAB [Hammo et al., 2004]		✓			✓				✓	
ArabiQA [Benajiba et al.,2007]		✓		✓				✓		✓
QASAL [Brini et al., 2009]			✓	✓				✓		
DefArabicQA [Trigui et al., 2010]		✓		✓				✓		
IDRAAQ [Abouenour et al.,2012]		✓				✓		✓		✓
AquASys [Bekhti & Al-Harbi, 2013]	✓				✓			✓		
ALQASIM [Ezzeldin et al., 2013]				✓					✓	✓
System of [NBdour and Gharaibeh, 2013]			✓				✓		✓	
JAWEB [Kurdi et al., 2014]				✓				✓		
Al-Bayan [Abdelnasser et al., 2014]		✓				✓		✓		

Ensuite, une performance globale de ces systèmes sera exposée par le tableau 2.10. Cette performance représente les résultats parvenus par plusieurs travaux en arabe qui sont

nettement mentionnés dans la (section 3.2). En outre, nous fournissons une explication approfondie de chaque recherche en fonction de leurs résultats expérimentaux. D'après les évaluations montrées par les chercheurs, nous abordons que les résultats parvenus par ces paradigmes sont encourageants. Enfin, tous les résultats et les explications sont résumés dans les tableaux 2.10 et 2.11:

Tableau 2.10: Etude comparative des systèmes de question-réponse en arabe (1/2)

Système / Critères	AQAS	QARAB	ArabiQA	QASAL	DefArabicQA	AquASys
Domaine ouvert / restreint	Restreint : radiation	Ouvert	Ouvert	Ouvert	Ouvert	Ouvert
Langage d'implémentation	Non mentionné	Non mentionné	Java	Non mentionné	Java	Non mentionné
L'appui sur WORDNET	-	-	-	-	-	-
L'appui sur l'ontologie	-	-	-	-	-	-
Ressources linguistiques	-	Abuleil's tagger [Abuleil & Evens, 2002],	-	Plate forme NOOJ	-	-
Approche	Utilise un modèle basé sur la connaissance. Recherche dans des bases de données structurées	Traite la question comme un "sac de mots". Le module de recherche d'information est basé sur le modèle d'espace vectoriel de Salton. Reconnaît les entités nommées	catégorise la question on (nom, date, la quantité, et la définition) selon les mots d'interrogation et attribue un rang plus élevé pour les passages qui ont une plus petite distance entre les mots-clés: densité de la distance.	QASAL utilise la plate-forme NooJ comme un environnement de développement linguistique pour répondre aux questions factuelles.	le système identifie les définitions candidates en utilisant un ensemble de les schémas lexicaux puis il filtre ces définitions en utilisant des règles heuristiques et les classe selon une approche statistique.	Le système segmente la question en nom interrogative, verbe de la question et des mots-clés.
Source	Des données structurées	Des données non structurées (corpus du journal Al-Raya)	Corpus	Livre d'éducation	Web	Corpus
Réponse	Phrase	Passage court	Phrase	Phrase	Phrase	Phrase courte
Type de question	Plusieurs formes (énoncés déclaratifs)	Question commencé par: من،متى، أين، كم	Question factuelle	Question factuelle	Question de définition	Question factuelle
Caractéristiques	Enlèvement des mots vides	Type et catégorie de réponse attendue.	Reconnaissance des Mots-	Type de réponse	Thème se la	Mots-clés et type de

	Tokenization		clés et entités nommées.	attendue, focus et mots clés.	question, le type de réponse attendue.	réponse attendue.
Performance	Non mentionné	Précision : 97.3% Rappel : 97.3 %	Précision : 83.3%	Non mentionné	MRR: 0.81	Précision : 66.25% Rappel : 97.5% F1-Score : 87.89%

Tableau 2.11: Etude comparative des systèmes de question-réponse en arabe (2/2)

Système / Critères	**IDRAAQ**	**ALQASIM**	**Système de NBdour et Gharaibeh**	**JAWEB**	**Al-Bayan**
Domaine ouvert / restreint	Ouvert	Ouvert	Ouvert	Ouvert	restreint : Coran
Langage d'implémentation	Java	Non mentionné	Non mentionné	Dreamweaver, Java	Non mentionné
L'appui sur WORDNET	WordNet arabe	WordNet arabe	-	-	-
L'appui sur l'ontologie	Ontologie SUMO Ontologie YAGO	-	-	-	Ontologie Coran
Ressources linguistiques	Analyseur Al-Khalil	MADA+TOKAN [Habash et al., 2009]	-	Arabic Khoja's stemmer	Outil LingPipe Lucene
Approche	IDRAAQ utilise une version enrichie de Word Net arabe	ALQASIM répond les questions à choix multiples de la collection de QA4MRE @ CLEF 2013. Il utilise une nouvelle technique en analysant les documents de test de lecture à la place des questions.	Le système est basé sur une approche de représentation logique sémantique pour récupérer le paragraphe qui contient des réponses pertinentes à la question.	Jaweb analyse les questions et extrait les informations importantes pour récupérer les réponses les plus pertinentes à partir d'un corpus arabe. Il fournit une interface d'utilisateur.	Al-Bayan Comprend la sémantique du Coran et répond aux questions des utilisateurs en utilisant le Coran et ses livres d'interprétation (tafsir).
Source	-	Collection de QA4MRE @ CLEF 2013	Corpus	Corpus	Coran et ses livres d'interprétation (tafsir)
Réponse	Phrase	Phrase	Paragraphe	Phrase	Phrase
Type de question	QCM	QCM	Question de type oui/non	Question factuelle	Question factuelle

Caractéristiques	Mots clés, type de réponse attendue.	Non mentionné	Mots vides	Type de réponse attendue, mots clés.	Reconnaissance des entités nommées, type de la question, type de réponse attendue
Performance	Exactitude : 0.13, C@1: 0.21 %	Précision : 0.31% C@1: 0.36 %	Non mentionné	Rappel : 15-20%	Précision: 85%

A partir de l'analyse précédente, nous pouvons confirmer que depuis l'introduction de la question-réponse arabe il n'y a presque pas de recherche qui ont été fait à l'aide de théorèmes et de raisonnement profond. De ce fait, répondre à des questions en utilisant des approches fondées sur la logique ou l'inférence reste un champ d'étude trop peu abordé dans la littérature de la question-réponse arabe. A notre connaissance, nous constatons que ces approches sont trop rares en question-réponse arabe. De plus, dans nos travaux de recherche, nous présentons un aperçu des principales approches et systèmes proposées en arabe, cette analyse est profitable pour de nouvelles orientations de recherche dans ce domaine. C'est utile d'examiner une analyse performante des différentes études proposées pour l'arabe vu la rareté des enquêtes dans ce domaine. Cette analysée est bien détaillée et décrite dans [Bakari et al., 2015] et [Bakari et al., 2016].

4.1 **Avantages et limites des systèmes proposés**

Nous avons bien présenté dans ce chapitre les différents systèmes de question-réponse en arabe. Chacun d'eux a ses avantages et ses limites. Comme nous pouvons le constater les approches adoptées par ces systèmes sont diverses. Nous pouvons également constater que ces différentes approches partagent souvent des points communs. De toute façon, nous ne pouvons pas dire qu'il n'existe pas un système plus efficace que l'autre. Aussi, la production d'un nouveau système n'écarte pas l'existence d'un ancien système. A vrai dire, les systèmes actuels donnent de meilleurs résultats. Néanmoins, malgré les avantages qu'offrent ces systèmes ils ont certaines limites. De plus, l'arabe est parmi les langues qui sont moins étudiées par les chercheurs dans le domaine de la question-réponse. Par conséquent, les limites et le manque de recherches ont engendré la motivation pour proposer une nouvelle approche et donner des pistes pour elle. Nous révélons les avantages des systèmes de question-réponse dans le tableau 2.12 :

Tableau 2.12: Avantages et limites des systèmes proposés

Système	Type de Question	Expérimentation	Contributions	Limites
AQAS [mohamed et al., 1993]	Enoncés déclaratifs	Aucune évaluation publiée n'est disponible pour le système	AQAS est considéré le premier prototype pour a question-réponse arabe.	L'analyse morphologique utilise un dictionnaire de taille limitée.
QARAB [Hammo et al., 2004]	Question commencée par : من،متى، أين، كم	Précision : 97.3% Rappel 97.3 %	QARAB se base sur une approche favorisant une liaison entre un système de recherche d'information et un système de TALN qui réalise l'analyse linguistique.	Le système fournit des réponses à des questions factuelles, mais ne supporte pas les autres types de questions.
ArabiQA [Benajiba et al., 2007]	Question factuelle	Précision : 83.3%	L'approche proposée se base sur une liaison entre un système de recherche d'information et un système de TALN qui réalise l'analyse linguistique.	L'implémentation du système n'a pas été achevée.
QASAL [Brini et al., 2009]	Question factuelle	Aucune évaluation publiée n'est disponible pour le système	QARAB se base sur une approche favorisant une liaison entre un système de recherche d'information et un système de TALN qui réalise l'analyse linguistique.	Pas de résultats expérimentaux ou des mesures de performances publiées. La fonctionnalité globale du système est limitée.
DefArabicQA [Trigui et al., 2010]	Question de définition	MRR: 0.81	Le premier système en arabe qui traite de la question de définition.	L'expérimentation contient que 50 organisation de question de définition et les réponses ont été évaluées par un seul locuteur natif arabe.
IDRAAQ [Abouenour et al.,2012]	QCM	Exactitude: 0.13, C@1: 0.21 %	Le système traite un type différent de questions (QCM) par rapport à la majorité des études en arabe qui traitent de questions factuelles.	Les expériences fournies par IDRAAQ identifient les lacunes du système lors du traitement des questions non factuelles et au stade de validation de la réponse.
AquASys [Bekhti & Al-Harbi, 2013]	Question factuelle	Précision : 66.25% Rappel : 97.5% F1-Score : 87.89%	Le système utilise intensivement les techniques de la TALN pour l'analyse de la question et récupère des réponses.	Les épreuves de test sont fondées sur un corpus non balisé
ALQASIM [Ezzeldin et al., 2013]	QMC	Précision : 0.31% C@1: 0.36 %	Le système traite un type différent de questions (QCM) par rapport à la majorité des études en arabe qui traitent de questions factuelles.	De nombreux mauvaises réponses aux questions sont responsables et de la liste des questions et des questions qui ont été mal traduits en raison de la traduction automatique erronée.
System of [NBdour and Gharaibeh, 2013]	Question oui/non	Non mentionné	Le premier système inclut le type de questions oui/non en arabe.	Les auteurs proposent une approche de représentation logique fondée sur la sémantique, mais il ne couvre pas tous les composants du système (seulement de l'analyse de la question).
JAWEB [Kurdi et al., 2014]	Question factuelle	Rappel: 15-20%	une approche à base du Web qui fournit une interface pour l'utilisateur.	Le système ne fonctionne que pour certains types de questions (questions factuelles), mais ne supporte pas les autres types de questions.

Al-Bayan [Abdelnasser et al., 2014]	Question factuelle	Précision: 85%	L'approche présentée construit un modèle de recherche d'information sémantique.	Les auteurs ont révisés manuellement les 1200 concepts et leurs versets.

4.2 **Tendances actuelles de la question-réponse en arabe**

Pour améliorer la performance des systèmes de question-réponse en arabe, il y a eu des progrès considérables en raison des efforts déployés par les chercheurs dans ce domaine. Initialement, la plupart des recherches dans le domaine de la question-réponse arabe ont porté sur des approches morpho-syntaxiques. Récemment, il y a eu des tendances qui ont été faits sur la sémantique, la logique, l'inférence et le raisonnement profond des approches proposées. A cet égard, les chercheurs ont tendance à effectuer une analyse profonde des textes succeptibles de contenir la réponse à une question en langage naturel. Ceci est effectué grâce à des approches plus profondes. D'ailleurs, des efforts ont proposé l'utilisation des approches sémantiques en intégrant des ontologies ou en contrôlant les vocabulaires pour améliorer la question-réponse. A titre d'exemple, [Abuenour, 2014] a utilisé WordNet arabe pour étendre la requête de l'utilisateur en capturant des termes qui sont liés de façon sémantique aux termes de l'utilisateur. D'autres systèmes ont tenté d'intégrer la connaissance du discours et d'autres techniques de TALN profondes.

Quelques autres travaux ont exploré l'analyse sémantique de textes arabes. Ils présentent l'information sémantique dans les graphes conceptuels qui sont un formalisme puissant et prometteur, spécialement avec une plate-forme conceptuelle basée sur le graphe comme Amine[13]. Ce formalise peut également être utile pour les systèmes de question-réponse en utilisant une analyse sémantique de la question et en la comparant à l'analyse sémantique des documents ou des passages candidats. En d'autres termes, les recherches dans le domaine de la question-réponse arabe sont en retard dans le développement de ressources du Web sémantique et l'exploitation de leur richesse pour le développement des systèmes. Par exemple le système IDRAAQ [Abouenour et al., 2012] utilise des techniques Web sémantiques et des outils pour l'expansion des requêtes. Par conséquence, le champ de la question-réponse arabe est un champ de recherche très actif. De nombreuses approches (morpho-syntaxiques, logiques, sémantiques, etc), qui commencent à atteindre le grand public, sont là pour illustrer l'importance des avancées accomplies... mais de nombreuses questions en question-réponse arabe restent cependant en suspens.

[13] http://amine-platform.sourceforge.net/

Particulièrement, en question-réponse arabe, il se trouve des aspects qui ont été moins traités. Ces aspects concernent l'utilisation de la sémantique, l'incorporation de la logique et des mécanismes de raisonnement et la détection de l'implication textuelle. Nous n'avons rencontré que quelques approches qui ont été tentées dans la sémantique. Certains travaux ont adapté les approches de question-réponse pour utiliser les ontologies arabes [AlAgha & Abu-Taha, 2015]. À notre connaissance, il existe peu de travaux qui fournissent des approches fondées sur la logique et l'inférence en langue arabe. Par exemple, [N Bdour et Gharaibeh, 2013] ont proposé un système de réponse à une question arabe basé sur la récupération des paragraphes ; ces auteurs ont utilisé un corpus de 20 documents arabes et une collection de 100 questions oui / non différentes qui se transforment en une représentation logique. Néanmoins, en ce qui concerne leur proposition, nous ne trouvons aucune information pour son séquençage. En outre, le manque ou l'absence de ce type d'approches en arabe suggère que la pertinence et la faisabilité d'explorer des approches basées sur la logique sémantique pour la réponse aux questions en arabe est primordiale.

4.3 **Positionnement de nos travaux de recherche**

Nos travaux mettent l'accent sur quelques dimensions dont le développement semble plus particulièrement à la compréhension automatique des textes arabes. Plus précisément, nous parvenons à introduire une réelle analyse de texte contenant la réponse à une question. Isolément du domaine étudié en question-réponse, la problématique de comprendre un texte et d'assurer son analyse en profondeur, n'ont pas possédé une priorité suffisante sauf dans les cinq dernières années en comparaison avec celle donnée à l'analyse de question. Nous suggérons que la compréhension automatique d'un texte figure parmi les composants de base d'un tel système de question-réponse pour choisir et obtenir une réponse précise. Par conséquent, l'approche mise en œuvre dans nos travaux de thèse permet de construire une représentation sémantique et logique des textes via les graphes conceptuels. De ce fait, la compréhension du texte peut être définie comme un processus cognitif de compréhension des concepts à partir d'un texte et des relations entre eux.

De nombreuses recherches ont été proposées pour l'analyse des textes en anglais, mais il existe peu de recherches pour l'analyse et la compréhension des textes arabes. Par conséquent, ce travail cherche à développer une nouvelle approche dont le but d'analyser des textes en utilisant une approche sémantique, où un texte est représenté sous forme de graphes conceptuels. En fait, la représentation du texte arabe sémantiquement à l'aide de graphes

conceptuels est l'une des techniques récentes qui facilitent le processus de manipulation du domaine de la question-réponse arabe. De plus, le fait de représenter le texte avec un graphe conceptuel peut faciliter le processus de génération des représentations logiques d'un texte arabe. Ainsi, notre approche a été mise en œuvre pour générer non seulement des graphes conceptuels et des représentations logiques à partir d'un texte et de la question, mais elle cherche à déterminer des implications textuelles entre ces représentations logiques.

Nous présentons en détail dans les chapitres suivants une approche sémantique et logique pour une modélisation conceptuelle et logique du contenu textuel et pour la détection de l'implication textuelle pour la question-réponse arabe. Notant que le manque de recherches dans ce domaine nous encourage à proposer notre approche qui repose sur une implication logique pour améliorer la recherche en question-réponse arabe.

Conclusion

Dans ce chapitre, nous avons étudié les systèmes de question-réponse surtout ceux proposés en arabe. En effet, nous avons commencé par une motivation pour un système de question-réponse en arabe. Nous avons vu également un aperçu sur la question-réponse dans certaines langues. Plus précisément, nous avons présenté un aperçu sur les approches et les systèmes proposés. En arabe, les systèmes sont fondés essentiellement sur des approches morphosyntaxiques. En fait, l'intégration des approches fondées sur la logique à la question-réponse est une nouvelle alternative. À l'heure actuelle, les approches fondées sur la sémantique et/ou la logique et l'implication textuelle sont peu étudiées en arabe.

Dans le chapitre suivant, nous allons étudier la construction de notre corpus de textes-questions et présenter les fondements thèoriques pour une nouvelle approche en arabe.

Ensuite, nous allons focaliser, dans le chapitre 4, sur la proposition d'une nouvelle approche pour la question-réponse arabe. Cette approche permet d'améliorer ce domaine de recherche par l'intégration de la sémantique et la logique et l'inclusion de la technique de l'implication textuelle.

CHAPITRE 3 : CONSTRUCTION DU CORPUS AQA-WEBCORP ET FONDEMENTS THEORIQUES POUR UNE NOUVELLE APPROCHE

Introduction……..…….... 71

1. L'analyse des questions en arabe.... 71

1.1 Analyse des questions pour les systèmes de question-réponse arabes.... 71

1.2 Apports de l'analyse de la question lors de l'extraction de la réponse 73

2. La construction d'un corpus.... 74

2.1 Utilisation du Web comme une source de corpus 74

2.2 Pourquoi introduire de nouveaux corpus pour l'arabe ?.... 75

3. Construction du corpus AQA-WebCorp.... 77

3.1 La collecte des questions.... 78

3.2 Présentation du corpus AQA-WebCorp 79

3.3 Démarche de la construction.... 81

3.4 Passages générés.... 83

4. La compréhension automatique de textes 84

4.1 Motivation pour la compréhension automatique de textes.... 84

4.2 Applications de la compréhension automatique de textes 85

4.3 Une représentation sémantique pour la compréhension 86

4.4 Une interprétation logique pour la compréhension 91

5. La reconnaissance d'implications textuelles.... 95

5.1 Motivation 96

5.2 Applications de l'implication textuelle.... 97

5.3 Implication textuelle en question-réponse 98

5.4 Traitement logique de l'implication textuelle 101

Conclusion…….…….... 102

Introduction

Dans la totalité des systèmes de question-réponse, la génération d'une réponse précise à une question en langue naturelle passe nécessairement par des analyses (de la question ou des passages). En effet, l'analyse de la question est une tâche importante voire nécessaire non seulement pour la recherche de documents mais aussi pour l'extraction d'une réponse justifiable et précise. Dans ce chapitre, nous exposons le terrain de notre approche, nous présentons ainsi l'analyse des questions en langue arabe, l'extraction des textes à partir du web ainsi que leur analyse. En effet, les deux méthodes d'analyse sont successivement proposées pour les questions et les passages. Ces deux méthodes sont basées essentiellement sur une représentation sémantique et logique. D'abord, nous étudions notre méthode de construction du corpus AQA-WebCorp. Nous illustrons à la fin de ce chapitre la compréhension automatique des textes arabes et l'implication textuelle.

1. L'analyse des questions en arabe

L'analyse des questions constitue une tâche primordiale dans le cadre des systèmes de question-réponse, car pour répondre proprement à une question, il faut d'abord bien l'analyser. En arabe, il y a des études menées pour obtenir des meilleurs résultats pour l'analyse des questions. Cette analyse varie d'une étude à une autre. En effet, la majorité de ces tentatives se concentrent sur la reconnaissance des entités nommées, sur la détermination de mots clés présentés dans la question, etc. Dans cette section, nous présentons un aperçu des différentes tâches qui sont couvertes par diverses études proposées en arabe à propos de l'analyse de la question.

1.1 Analyse des questions pour les systèmes de question-réponse arabes

Plusieurs travaux indiquent que la tâche d'extraire une réponse à une question donnée nécessite essentiellement une analyse profonde de la question [Embarek, 2008]. Cette analyse a extrait les principales caractéristiques de la question [Zweigenbaum et al., 2008], à savoir le type de la réponse attendue, l'objet de la question (focus), les termes qui seront utilisés dans la recherche des documents. Ces caractéristiques seront prises en compte dans les prochaines étapes de la recherche des réponses précises [Rodrigo et al., 2010]. Ainsi, un autre objectif complémentaire et essentiel de cette étape est de déterminer les entités nommées dans les questions d'entrée et d'aborder les relations qui les relient.

A cet égard, nous présentons dans notre étude une analyse de performance des différentes recherches en arabe [Bakari et al., 2015]. Nous introduisons régulièrement une analyse des principales tâches pour un système de question-réponse (analyse de la question, la récupération des passages et l'extraction de la réponse, etc.). Dans notre proposition, l'analyse des questions pourrait déterminer le type de la réponse attendue et les mots clés et elle reformule la question en une forme déclarative afin de générer des représentations sémantiques et logiques. Les détails des caractéristiques étudiées par ces études sont donnés dans le tableau 3.13.

Tableau 3.13: Description de l'analyse des questions: enquêtes arabes

Système	Description des tâches
QARAB [Hammo et al., 2004]	Extrait le type et la catégorie de la réponse souhaitée (nom, lieu, quantité ...).
Work of [Rosso et al. 2006]	Extraire les mots-clés des questions; reconnaître la question entités nommées des questions, classer les questions.
ARABIQA [Benajiba et al., 2007]	Classer la question; extraire les mots clés et les entités nommées
QASAL [Brini et al.,2009]	Formuler la requête; extraire le type de réponse attendu, l'accent de la question et les mots clés de la question.
Work of [Kanaan et al., 2009]	Segmenter la question (en entités); déterminer le type et l'accent (expression nominale appropriée), extraire la racine de tous les mots vides.
DefArabicQA [Trigui et al., 2010]	Identifier le sujet de la question (c'est-à-dire NE) et dédier le type de réponse attendu
IDRAAQ [Abouenour et al. 2012]	Extraire les mots clés; reconnaître le type de réponse attendu; créer la requête.
AQuASys [Bekhti & Al-Harbi, 2013]	Identifier le type de réponse attendu; segmenter la question en nom interrogatif, verbe de la question et mots clés de la question.
Système de [N Bdour and Gharaibeh, 2013]	Supprimer le point d'interrogation et la particule interrogative; segmenter (en entités); supprimer les mots vides et les particules de négation, étiqueter, analyser.
JAWEB [Kurdi et al., 2014]	Trouver les jetons; détecter le type de la réponse; extraire les mots clés de la question; générer le mot clé supplémentaire; extraire le « stem » du mot clé de la question.
Al-Bayan [Abdelnasser et al., 2014]	Classer les questions avec Support Vector Machine; extraire le type de la question, le type de réponse attendu et les entités nommées.

En effet, en observant le tableau 3.14, nous constatons que la plupart des systèmes dédiés pour cette langue portent sur la classification de la question. Cependant, dans nos travaux, nous proposons de générer des représentations sémantiques et logiques à partir des

questions. Cette étape est importante avant d'introduire une implication textuelle efficace. Cette implication pourrait nous aider à extraire la réponse précise.

Tableau 3.14: Tâches d'analyse des questions couvertes par les études arabes

	Traitement de la question		
Tâches d'analyse de la question / Système de question-réponse	Segmentation de la question	Classification de la question	Formulation de la question
AQAS [mohamed et al.,1993]			
QARAB [Hammo et al., 2004]		✓	
ArabiQA [Benajiba et al.,2007]		✓	
QASAL [Brini et al., 2009]			✓
DefArabicQA [Trigui et al., 2010]		✓	
IDRAAQ [Abouenour et al., 2012]		✓	
AquASys [Bekhti & Al-Harbi, 2013]	✓		
ALQASIM [Ezzeldin et al., 2013]			
System of [NBdour and Gharaibeh, 2013]			✓
JAWEB [Kurdi et al., 2014]			
Al-Bayan [Abdelnasser et al., 2014]		✓	

En outre, dans d'autres cas, l'analyse des questions est basée sur une réécriture de cette question afin d'accroître la capacité des autres modules à identifier la réponse. D'autres études portent particulièrement sur la classification des questions comme le premier problème de la recherche en question-réponse [Burger et al., 2001]. Selon [Dhanjal & Sharma, 2015], la classification des questions possède un rôle essentiel dans le système de question-réponse en identifiant le type de la question et le type de réponse attendue. Cette classification présente certaines informations sur la question. En effet, elle implique des traitements sur la question pour identifier la catégorie de réponses que l'utilisateur veut trouver. Cette étape est accomplie en tirant profit des informations obtenues à partir de la segmentation de la phrase. Cette segmentation extrait les noms, les verbes, les adjectifs et les prépositions [Moreale & Vargas-Vera, 2004].

1.2 Apports de l'analyse de la question lors de l'extraction de la réponse

En contexte de la question-réponse, le rôle de l'analyse de la question pour extraire la réponse correcte est un sujet de nombreuses études [Zwaigenbaum et al., 2008], [Mendes & Véronique, 2004]. En effet, le but principal d'une telle analyse est de définir l'information en

question qui pourrait de faciliter la tâche de génération de réponses en langage naturel. Dans cet esprit, Zweigenbaum et ses associés ont montré dans leur travail que le recours à une analyse efficace de la question est de plus en plus fréquent pour trouver la réponse précise. En effet, l'analyse de la question comprend la capacité de déterminer la forme de la question. Quand la réponse est factuelle, il suffit de spécifier le type de la réponse attendue [Zwaigenbaum et al., 2008]. D'autres chercheurs recommandent que certains éléments de l'analyse de la question, à savoir, sa structure syntaxique, le type de réponse attendue, la lexicalisation des concepts, etc., puissent contribuer à générer de bonnes qualités de réponses linguistiques [Mendes & Véronique, 2004]. De plus, d'autres chercheurs comme Brill et ses collègues suggèrent que la contribution des différents composants influe à l'exactitude globale du système de questions-réponses [Brill et al., 2002].

2. La construction d'un corpus

Un corpus est une collection de morceaux de textes sous une forme électronique, sélectionnés selon des critères externes pour représenter, autant que possible, une langue en tant que source de données pour une recherche linguistique [Sinclair, 2005]. En effet, une définition qui est à la fois précise et générique d'un corpus, selon [Rastier, 2005], est le fruit des choix des linguistes qui le rassemblent. Un corpus n'est pas un objet simple, ne doit pas être une simple collection de phrases ou un « sac de mots ». C'est en fait un assemblage de textes. Cet assemblage peut couvrir de nombreux types de textes. En outre, la construction de corpus est généralement utilisée pour de nombreuses applications de TALN, y compris, la traduction automatique, la recherche d'information, la question-réponse, etc. Dans ce sens, plusieurs tentatives de travaux ont réussi à définir un corpus. Ainsi, la construction d'un corpus est une tâche qui est à la fois primordiale et délicate. C'est une tâche complexe car elle dépend en large partie d'un nombre important de ressources à exploiter. Une façon de diminuer ce problème est d'utiliser le Web comme source de données. En effet, le Web est une quantité colossale de textes a été récupéré librement [Gatto, 2011]. Il contient des milliards de mots de texte qui peuvent être utilisés pour tout type de recherche linguistique [Kilgarriff & Grefenstette, 2001].

2.1 Utilisation du Web comme une source de corpus

Ces jours, le Web joue un rôle influent pour la recherche d'information. Il est considéré la plus grande ressource de connaissances (textuelles, graphiques ou sonores). Ces

ressources adoptent la construction de corpus. Mais leur constitution n'est pas facile de telle sorte qu'elle soulève un certain nombre d'interrogations [Isaac et al., 2001]. C'est une source immense, gratuite et disponible. Maintenant le web est avec nous, avec un simple clic du bouton, des quantités colossales de textes ont été récupérées gratuitement [Gatto, 2011]. En effet, les demandes des utilisateurs pour des informations précises est toujours en augmentation, en particulier, avec l'expansion de l'information numérique. Cette croissance est non seulement consacrée à consulter les documents existants sur le Web, mais aussi à construire des corpus pour plusieurs applications du langage naturel. Néanmoins, la construction d'un corpus de textes à partir du web n'était pas une tâche simple. Une telle constitution a contribué au développement et à l'amélioration de plusieurs outils linguistiques, tels que les systèmes de question-réponse, les systèmes d'extraction d'information, les systèmes de traduction automatique, etc.

Initialement, pour chercher des données du Web, des moteurs de recherche sont utilisés afin de donner une liste de documents en réponse à une sélection de mots réalisée par l'utilisateur. La liste de documents est censée de contenir les documents les plus pertinents par rapport à la sélection de mots. De ce fait, la pertinence est mesurée en suivant l'existence des mots dans ces documents, la quantité, la visibilité des documents dans la collection, les caractéristiques saillantes. D'ailleurs, la tâche de construire un corpus de ressources textuelles à partir du Web est un peu récente. D'où, la collecte, l'organisation et l'utilisation des ressources à partir du web est difficile. En outre, le Web a également été utilisé par des groupes à Sheffield et Microsoft comme étant une source de réponses pour les applications de question-réponse, dans une fusion de moteur de recherche et des technologies de traitement du langage naturel ([Susan et al, 2002], [Mark et al., 2002]). AnswerBus permet de répondre aux questions posées en anglais, allemand, français, espagnol, italien et portugais [Zhiping, 2002].

2.2 **Pourquoi introduire de nouveaux corpus pour l'arabe ?**

Bien que, l'arabe est tenu en compte parmi les dix premières langues sur le Web, il manque de nombreux outils et ressources. Ainsi, la langue arabe est la langue officielle dans toutes les nations arabes comme la Tunisie, l'Egypte, l'Arabie Saoudite et l'Algérie. En outre, c'est aussi une langue officielle dans les pays non arabes comme le Tchad et l'Erythrée. Malheureusement, il y a quelques attentions accordées aux corpus arabes, aux dictionnaires lisibles par machine, aux lexiques [Hammo et al., 2004]. Par conséquent, il est si difficile de

trouver un corpus dédié pour le traitement de la langue naturelle et plus spécifiquement pour la langue arabe. Ainsi, le manque et / ou l'absence de corpus en arabe ont été un problème pour la mise en œuvre du traitement du langage naturel. Afin de mieux se situer dans ce cadre, nous décidons de construire un corpus de questions et de textes recueillis du WEB qui a été bien décrit dans notre travail [Bakari et al., 2016].

Ainsi, dans leur travail [Bekhti & Al-Harbi, 2013], les chercheurs présentent que les systèmes de question-réponse arabes développés sont encore peu nombreux par rapport à ceux développés, par exemple, pour l'anglais ou le français. Ceci s'explique essentiellement par deux raisons: le manque d'accessibilité aux ressources et aux outils linguistiques, tels que les corpus et les outils arabes du TALN, et la nature très complexe de la langue elle-même (la langue arabe est infléxionnelle et non concaténée et il n'y a pas de capitalisation comme dans le cas de l'anglais). De leur côté, dans leur enquête [Ezzeldin & Shaheen, 2012], les auteurs signalent que comme toute autre langue le traitement du langage naturel arabe nécessite des ressources linguistiques, telles que des lexiques, des corpus, des banques d'arbres et des ontologies qui sont essentielles pour les tâches sémantiques avec l'apprentissage automatique et syntaxiques ou la recherche et la validation des mots traités. Par conséquent, nous constatons que peu sont les études qui sont fondées sur leurs propres corpus, notamment en question-réponse arabe.

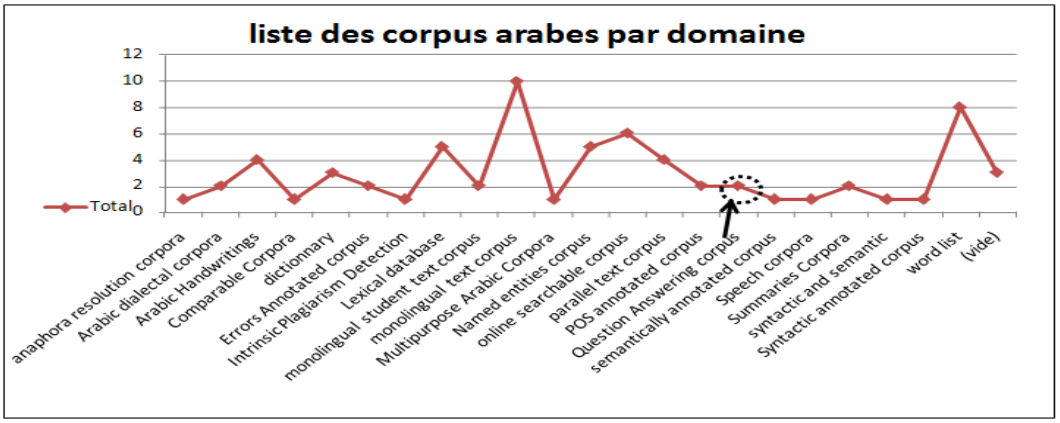


Figure 3.5: Liste des corpus arabe dans plusieurs domaines par rapport à la question-réponse

Malgré que les efforts de construction des corpus de textes sont concentrés sur l'anglais. Les corpus arabes peuvent également être acquis à partir du Web comme grande source de données. Ces tentatives pourraient s'avérer dans la totalité des applications de TALN. Cependant, nous remarquons aussi des efforts non négligeables, essentiellement pour la question-réponse. Ainsi, la plupart des études dans la construction de corpus arabes sont

conçues pour des domaines autres que la question-réponse (voir figure 3.5). D'après une recherche effectuée à Google[14] , nous constatons que le nombre de ses des tentatives est limité. C'est pourquoi nous avons décidé de construire notre propre corpus. Pour aborder cet objectif, il nous faut comme une étape intermédiaire l'interrogation d'un moteur de recherche. La construction de corpus pour la question-réponse arabe s'améliore. Nous souhaitons bien qu'elle continuera à s'améliorer dans les prochaines années et achèvera un jour par produire des corpus dans ce domaine de recherche qu'ils devront être utilisés par les chercheurs dans leurs expérimentions.

A cet égard, dans le major de nos connaissances, le nombre de corpus dédié pour la question-réponse arabe est un peu limité. Parmi les études qui ont dédié pour le domaine de la question-réponse, nous citons [Trigui et al., 2010] qui a construit un corpus pour les questions de définition arabes traitant les définitions des organisations. Ces auteurs utilisent une série de 50 questions de définition de l'organisation. Ils ont expérimenté leur système en utilisant 2000 extraits retournés par la version de Wikipedia arabe et le moteur de recherche Google.

3. **Méthode proposée pour la construction du corpus AQA-WebCorp**

Avec le développement des médias électroniques et l'hétérogénéité des informations arabes sur la toile, l'idée de construire un corpus propre pour certaines applications du traitement du langage naturel, notamment la traduction automatique, la recherche d'information, la question réponse, est devenue de plus en plus pressante. Dans nos travaux, nous cherchons à créer et développer notre propre corpus de paires de questions-textes. Cette constitution permettra alors d'offrir une meilleure base pour notre expérimentation. Nous cherchons également à modéliser cette constitution par une méthode pour l'arabe dans la mesure où elle récupère des textes à partir du web qui pourraient s'avérer être des réponses aux questions des utilisateurs. Pour le faire, il fallait développer un script java qui permet d'extraire à partir d'une question donnée des pages html. Puis, nettoyer ces pages dans la mesure d'avoir une base textuelle idéale et un corpus de textes bruts. Certaines investigations pour la construction des corpus arabes sont brièvement présentées par la suite de cette section.

Au cours des dernières années, de nombreux travaux de recherche ont été menés pour la tâche de construction de corpus d'études. La majorité de ces investigations reposent sur des approches statistiques. Ainsi, un des avantages des corpus est qu'ils peuvent exhiber aisément

[14] http://www.qatar.cmu.edu/~wajdiz/corpora.html

des données quantitatives qui ne peuvent pas fournir des intuitions de manière fiable. Nous allons suivre une démarche théorique validée par une investigation empirique pour fournir un meilleur corpus d'étude. Cette démarche fournit une compréhension d'une nouvelle approche via une revue de la littérature bien établie. Dans nos travaux de recherche, nous nous sommes intéressés à construire notre corpus de questions-textes AQA-WebCorp (Arabic Question-Answering Web Corpus) en interrogeant le moteur de recherche Google. En fait, Google a travaillé sur plusieurs initiatives pour aider à augmenter le contenu en langue arabe.

3.1 **La collecte des questions**

Nous avons recueilli les questions qui peuvent être posées dans différents domaines, y compris, le sport, l'histoire et l'islam, les découvertes et la culture, les nouvelles du monde, la santé et la médecine. En effet, la collecte de ces questions est réalisée à partir de plusieurs sources, à savoir, des forums de discussion, les questions communément posées (FAQ), quelques questions traduites à partir les deux campagnes d'évaluation TREC et CLEF (figure 3.6).

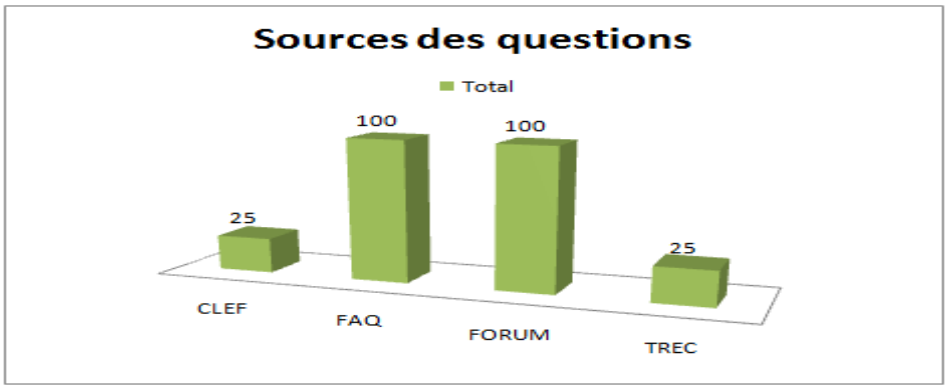


Figure 3.6: Source des questions utilisées dans notre corpus

La taille de notre corpus est dans l'ordre de 250 questions factuelles: 25 questions traduites de TREC, 25 questions traduites de CLEF et 200 questions recueillies à partir des forums et FAQ. En effet, le terme factuel (terme français inspiré du terme anglais factoid) : les questions factuelles se rapportent à un fait simple [Jurafsky & Martin, 2008] et attendent une entité nommée comme réponse [Forner et al., 2008]. Par conséquent, pour construire notre corpus, nous avons utilisé les textes arabes disponibles sur Internet qui sont recueillis en se basant sur les questions posées au départ.

Nous avons collecté 250 questions factuelles qui peuvent être de cinq catégories: soit une personne : من صمم برج ايفل ؟ (Who designed the Eiffel Tower?), soit une location: أين تقع شلاّلآت نيآغرآ ؟ (Where is the Niagara Falls ?), soit une date : متى استقلت تونس ؟ (When Tunisia became independent?), soit une organisation : ماهي عاصمة ماليزيا ؟ (What is the capital of Malaysia?) ou bien une expression numérique : كم يبلغ طول نهر الأمازون ؟ (How much is the length of the Amazon River?). Par conséquent, nous avons collecté des questions de type : (What, Where, When, Who, How) (ما، أين، متى، من كم) (figure 3.7).

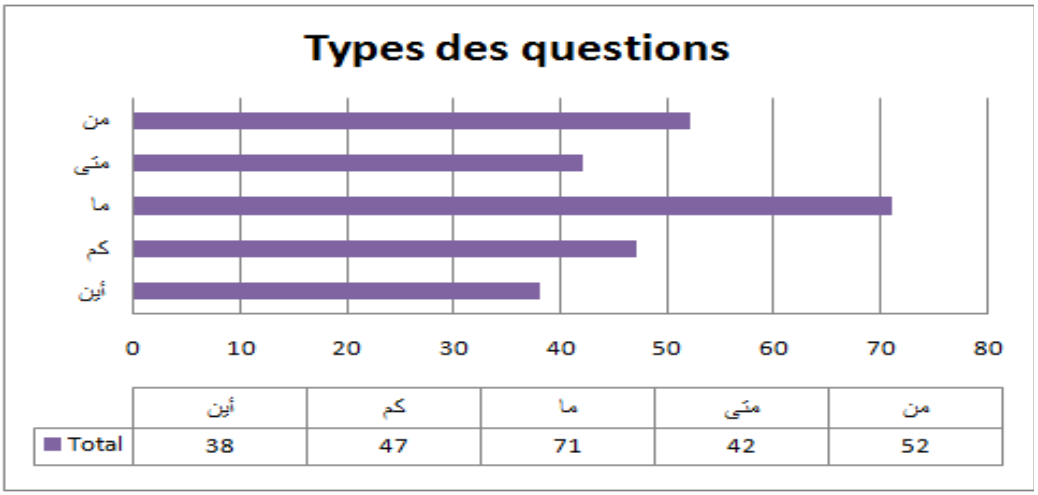


Figure 3.7: Types de questions

Après avoir collecté un ensemble de questions, la recherche de la réponse comporte plusieurs étapes, y compris, l'analyse de la question, ceci est une étape tout aussi importante qui est précieuse pour l'extraction de la réponse exacte. En particulier, nous nous attacherons d'une manière générale la façon dont nous analysons nos questions. L'enjeu que ce module cherche à soulever est: Quelle est la question?, Quel est le sujet de cette question?, Qu'est-ce que la question veut dire? En effet, nous analysons la question afin d'illustrer sa signification. Le chapitre 4 qui présente en détails notre approche qui montre la façon dont nous analysons les questions recueillies afin de générer leurs réponses. Toutes les caractéristiques obtenues par ce module sont données aux étapes suivantes pour notre système. Dans ce quit suit, nous détaillons les diverses étapes de notre méthode de construction du corpus AQA-WebCorp.

3.2 **Présentation du corpus AQA-WebCorp**

Nous montrons en détail la méthode proposée pour constituer notre corpus de paires de questions-textes, AQA-WebCorp dont son abréviation est (Arabic Question Answering Web Corpus). Ce corpus est construit par des textes recueillis à partir du web. Dans notre cas,

la taille du corpus obtenu dépend surtout du nombre de questions posées et du nombre de documents trouvés pour chaque question.

Dans le cadre de construction d'un corpus de textes à partir du web, les chercheurs dans [Issac et al., 2001] soulignent qu'il y a deux manières pour récupérer des informations à partir de la Toile pour construire un tel corpus. La première consiste à regrouper les données situées sur des sites connus [Resnik, 1998]. En effet, cette façon s'occupe d'un aspirateur à Web. Ce dernier assure la récupération des pages à partir d'une adresse donnée. En revanche, la seconde méthode interroge un moteur de recherche afin de sélectionner les adresses à partir d'une ou plusieurs requêtes (dont la complexité dépend du moteur). Puis, elle récupère soit manuellement soit automatiquement les pages correspondantes à partir de ces adresses.

Dans notre travail, nous suivons la deuxième méthode. Plus particulièrement, en se référant sur une liste de questions posées en langue naturelle, nous nous portons sur la récupération de la liste des URLS correspondantes. Puis, à partir de ces URLS, nous proposons de récupérer les pages web en relation. Enfin, nous proposons de nettoyer ces pages pour obtenir la liste de textes qui ont construit notre corpus. Après avoir construit notre corpus de paires de questions-textes, nous ne le conservons pas à cet état. Des étapes de post traitements et d'analyse seront bien occupées par la suite pour atteindre notre but de base qui est l'extraction de la réponse adéquate et précise à une question posée.

Dans une perspective d'analyse et de post traitement réalisée pour notre corpus en prenant en compte le contenu et la forme des questions, nous avons collecté des questions factuelles de type : (What, Where, When, Who, How) (ماء، أين، متى، من كم) (voir tableau 3.15). À cet égard, notre outil de construction de corpus est une liaison entre la demande de l'utilisateur et Google.

Tableau 3.15: Exemples de questions utilisées dans la construction de notre corpus

Types de questions		Exemples
What	ما(maa)	ماهي العملة المتداولة بتونس ؟ (ma hia al3mla almtadawla btounis)
Where	أين (ayna)	أين يقع جبل إيفرست؟ (ayna ya9a3 jabal 2iferset)
When	متى (mata)	متى استقلت تونس؟ (mata ista9alat tounis)
Who	من (man)	من أنت؟ (man anta?)
How	كم (kam)	كم طالبا في الجامعة؟ (kam Taaliban fil-jaami3a?)

3.3 **Démarche de la construction**

Pour mettre en ouvre notre corpus pour l'arabe, nous proposons une méthode robuste et simple implémentée en Java. Le principe de cette méthode est basé sur quatre étapes, relativement indépendantes. La constitution de notre corpus de paires de questions-textes en arabe se fait réellement en élaborant la totalité de ces quatre étapes. Cette méthode a fait l'objet d'un travail récemment publié [Bakari et al., 2016]. Nous décrivons dans ce qui suit chacune de ces étapes.

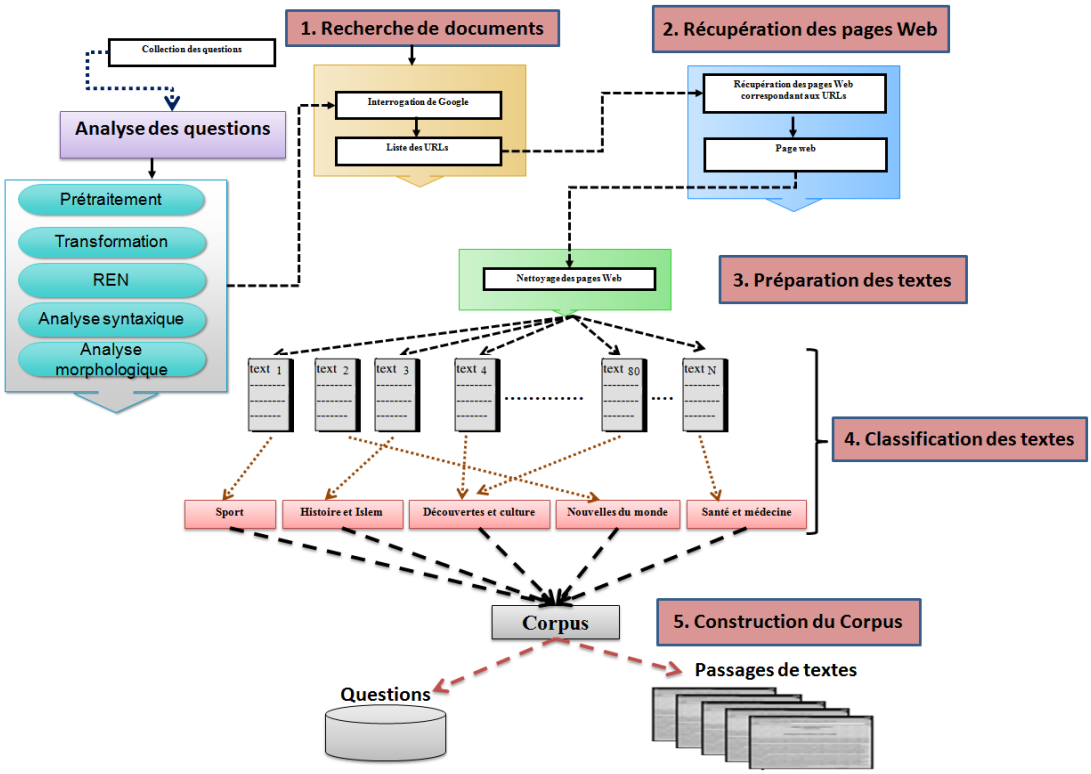


Figure 3.8: Processus de recherche de documents et d'extraction de passages pertinents

Comme illustré dans la figure 3.8, nous introduisons dans cette section une méthode simple de construction de notre corpus en favorisant une réelle interrogation du Web. Cette méthode est généralement construite de trois modules. Etant donné un module de génération d'une liste d'adresses URLs est mis en œuvre, ce module donne pour toute question posée en langue naturelle une liste d'URLs correspondantes. Ensuite, un module de transformation qui se comporte comme un générateur de pages Web correspondantes. Un troisième module assure une sorte de filtrage de ces pages. Le résultat de ce module pourrait être un ensemble de textes qui sont ajoutés aux questions pour construire notre corpus à partir du web.

(a) Recherche de documents

Dans notre cas, pour chercher une réponse à une question en arabe, nous proposons d'utiliser un moteur de recherche (par exemple, Google) pour récupérer les documents relatifs à chaque question. Nous pensons que c'est une meilleure solution, mais beaucoup plus complexe est la mise en œuvre d'un moteur de recherche linguistique pour un but particulier. En fait, la recherche de documents pertinents constitue l'une des étapes les plus pertinentes dans un système de question-réponse. À cet égard, l'interrogation d'un moteur de recherche accélère la récupération des documents mais nécessite un traitement hors ligne de ces documents. Après avoir récupéré les documents les plus pertinents pour chaque question, nous suggérons d'ajouter des post traitements linguistiques à ces documents qui sont en fait constituent notre corpus et peuvent avoir une réponse précise et appropriée.

(b) Récupération des pages web

Pour chaque question, nous proposons de rechercher la liste des adresses URLs qui correspondent à ces caractéristiques extraites dans l'étape de son analyse. En effet, le moyen d'accès au Web par défaut est un moteur de recherche tel que Google. Puis, pour chaque adresse donnée, nous récupérons la page web convenable. Le résultat de cette étape sera un ensemble de pages web. Plus particulièrement, il suffit de trouver la page HTML correspondante pour chaque URL donnée. Enfin, pour chaque page, un ensemble de passages de textes sera généré.

(c) Préparation des textes

Le but de cette étape est de transformer chaque page web obtenue à l'étape précédente en format ".txt". Les textes initialement sont en format ".html". Étant donné que l'application prévue est la modélisation statistique du langage, il semble justifié de les mettre dans le format ".txt". Pour cela, nous supprimons toutes les balises HTML pour chaque page extraite. Comme nous l'avons dit précédemment, notre méthode cherche des réponses à chaque question dans chaque texte généré. Il est possible soit de conserver le texte pour la construction de corpus, soit de l'ignorer.

(d) Classification de textes

Réellement, la classification de texte a été réalisée manuellement selon le sujet du texte et l'objet de la question. En outre, notre corpus est dédié à la question-réponse arabe. La taille de ce corpus est de l'ordre de 250 paires de questions et de textes (figure 3.9). Ceci a été construit en utilisant le web comme une source de données. Les informations collectées du web telles que les questions et les textes ont nous aidé à construire un corpus extensible. Ces données sont réparties sur cinq domaines « أخبار العالم , التاريخ والإسلام , إكتشافات و ثقافة , رياضة , صحة و طب».

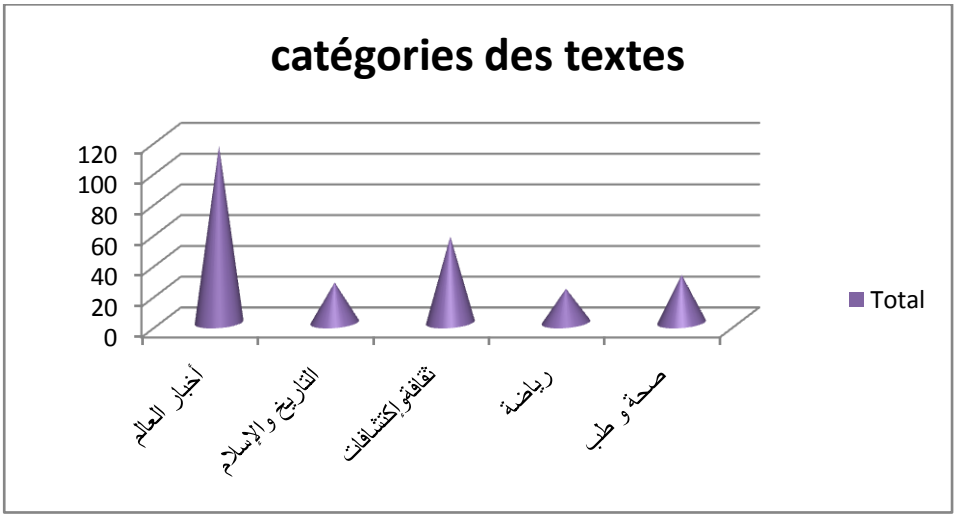


Figure 3.9: Statistiques des catégories de textes utilisés dans notre corpus

3.4 **Passages générés**

Avec un corpus, la recherche qualitative et quantitative linguistique peut être effectuée en quelques secondes, ceci permet d'économiser le temps et les efforts. Enfin, l'analyse des données de façon empirique peut aider les chercheurs non seulement de procéder à de nouvelles recherches linguistiques efficaces, mais aussi à tester les théories existantes. Initialement, nous avons identifié plusieurs éléments de l'analyse des questions qui pourrait favoriser la génération de réponses. Notre corpus est actuellement composé de 250 paires de questions-textes. Il est ainsi possible d'appliquer sur ces textes, préalablement à leur utilisation, un ensemble de traitements visant à rendre les processus suivants plus rapides ou à les normaliser. Nous espérons que nous pourrions continuer à faire progresser la construction de notre corpus, de sorte qu'il pourrait être efficacement utilisé à des objectifs diverses.

4. La compréhension automatique de textes

Le terme « compréhension automatique de textes » a vu le jour depuis le début du traitement automatique des langages dans les années 60-70. Ce concept est l'objet de nombreuses recherches et vise à saisir le sens global d'un texte. Les échecs récurrents des systèmes alors développés mettent rapidement en cause une vision trop générique de la compréhension automatique. En effet, de tels outils s'avèrent inutilisables dans un contexte opérationnel en raison du coût élevé des adaptations nécessaires (bases de connaissances et ressources lexicales spécifiques). Conscients d'être trop ambitieux au regard des possibilités technologiques, les chercheurs s'orientent alors vers des techniques plus réalistes d'extraction d'informations. S'il n'est pas directement possible de comprendre automatiquement un texte, le repérage et l'extraction des principaux éléments de sens apparaissent comme un objectif plus raisonnable. Cette réorientation théorique est reprise de façon détaillée par [Poibeau, 2003]. Le processus de compréhension proposé dans nos travaux est réalisé de deux façons. Une façon de représenter le texte en des graphes conceptuels et une deuxième façon de transformer ces graphes en des formes logiques de premier ordre.

[Patil et al., 2016], un texte est une série de phrases, où la phrase à son tour est une séquence de mots et de ponctuations qui sont combinées pour ajouter la sémantique au texte. En outre, un mot est une suite de caractères. Les textes construisent la masse d'information la plus présente sur le web (le son et les images sont plus récents). De ce fait, un texte est généralement composé par des phrases en taille équivalente avant leur extraction. Une raison est de décomposer un texte donné en une liste de phrases tant au niveau de la recherche d'une réponse précise à une question donnée. Une autre raison est que le texte en totalité peut ne pas engendre une réponse fiable aux questions des utilisateurs. Ainsi, pour répondre à ces questions, nous proposons ne pas tenir compte d'un texte long et d'aborder de nombreuses phrases de ce texte qui peuvent présenter la réponse désirée.

4.1 Motivation pour la compréhension automatique de textes

Comprendre un texte est tout à fait différent qu'effectuer une simple lecture. C'est la tâche d'atteindre une compréhension en profondeur d'un seul ou d'un petit nombre de textes. En fait, la tâche sera axée sur la lecture des documents simples où les bonnes réponses nécessitent une certaine inférence et un examen des connaissances de base acquises

auparavant [Banerjee et al., 2013]. C'est en fait, une manière où la compréhension humaine de texte a été mesurée pour répondre à des questions relatives au texte [Vanderwende, 2007].

Dans ce cadre, des chercheurs, tels que [Gomez-Adorno et al., 2013] définissent la compréhension automatique par la capacité de comprendre et de lire les principales idées écrites implicitement dans un texte donné. La réponse à une question à partir d'un texte donné pour évaluer la compréhension de ce texte est en fait une tâche très difficile. Cette dernière a été abordée dans plusieurs applications de TALN à savoir la question-réponse, la recherche d'information, etc. Le recours à cette technologie de question-réponse exige que le résultat doive être la réponse correcte à cette question, à la place d'un certain nombre de références à des documents qui contiennent la réponse. Ainsi, [Clark et., 2012] mentionnent que la compréhension automatique reste une tâche difficile. Elle demeure un défi majeur de l'intelligence artificielle. Son principal objectif est non seulement de comprendre un texte, mais aussi de construire un modèle interne cohérent du monde que le texte décrit.

En d'autres langues, la majorité des systèmes de question-réponse qui favorisant la compréhension automatique des textes utilisent essentiellement des prétraitements, à savoir, la résolution d'anaphores et de coréférences, la reconnaissance d'entités nommées, etc. Le système ayant obtenu des meilleurs résultats, pour l'anglais, à la tâche générale de QA4MRE 2011 et 2012 est celui développé par [Pakray et al., 2011], [Bhaskar et al., 2012]. Ce système utilise plusieurs mesures d'implications textuelles telles que comparaison des entités nommées, n-grammes et skip n-grammes communs et compare le type de la réponse avec celui attendu.

4.2 **Applications de la compréhension automatique de textes**

Diverses applications de TALN ont besoin de "comprendre" le sens d'un texte (même si elles le déterminent souvent de manière superficielle). La compréhension correspond au typage, à la reconnaissance et à la relation des éléments pertinents par rapport à une tâche. En outre, ces applications font intervenir un traitement sémantique de la langue naturelle. Dans certains cas, ces applications demandent une compréhension minimale du texte. L'annotation sémantique met en évidence au même texte des éléments d'information pertinents (mots clés, noms propres, etc.). L'extraction d'information permet d'extraire des informations structurées pour remplir une base de données (e.x. concernant des rachats d'entreprises, des réseaux d'interactions géniques, etc.). Le résumé automatique de textes vise à fournir un texte cible

résumant les principales informations contenues dans un (ou des) texte(s) source (s). La cible est généralement une simple sélection de phrases pertinentes du (ou des) texte(s) source (s). En question-réponse, c'est en réalité construire une représentation cohérente du contenu d'un texte pour répondre à une question en langue naturelle.

4.3 **Une représentation sémantique pour la compréhension automatique de textes**

La compréhension de textes est destinée à rendre compte de la signification du texte. En effet, comprendre un texte suppose que nous sommes capables de former une représentation cohésive et unifiée des informations émises par le texte. Cette compréhension requiert des analyses profondes et efficaces. Même pour l'analyse de textes, la modélisation de connaissances générales essentielles pour la compréhension reste un travail qui n'a été effectué qu'à une échelle très réduite, pour des applications très ciblées.

Les années 1980-1990 ont vu un fort déclin des systèmes de compréhension de textes proprement dits (au sens où il s'agit de donner une représentation d'un texte de manière globale) au profit des systèmes visant une compréhension très partielle, mettant en évidence quelques éléments essentiels et les relations qu'ils entretiennent entre eux. L'application à laquelle nous nous intéressons, et qui a été présentée dans nos travaux (un système de question-réponse), demande une compréhension du texte. Nous proposons de fournir une représentation sémantique de texte si nous voulons déterminer les implications nécessaires pour répondre à une question donnée. La tâche revient donc, dans une certaine mesure, à gérer la variation linguistique afin de fournir une représentation du texte qui permet de le manipuler.

Par ailleurs, le principal défi est de savoir comment convertir les informations trouvées dans le texte dans la langue avec laquelle une machine peut penser et prendre des décisions, et répondre correctement à des questions en langage naturel d'un utilisateur. En effet, notre objectif est de permettre une analyse approfondie des documents arabes (textes et questions) qui peuvent aller jusqu'à une représentation sémantique. Pour le faire, nous suggérons de recourir au formalisme des graphes conceptuels, qui permet une représentation plus riche du contenu textuel. En revanche, l'analyse de texte ne prétend pas remplacer l'interprétation de la signification des textes, il s'agit d'extraire des contenus ou une structure pour répondre à des questions précises. Ce formalisme représente la phrase du texte et la question avec une

structure formée par des sommets et des arêtes [Sowa, 1984]. Il fournit un niveau supérieur de compréhension du texte en capturant la sémantique dans le texte.

D'ailleurs, l'utilisation d'un graphe conceptuel pour représenter du texte a une très longue histoire dans le traitement de la langue naturelle, elle s'est concentrée sur les techniques de compréhension de la langue. Ainsi, l'analyse de textes devient de plus en plus importante dans les domaines de TALN, y compris, la question-réponse. Dans ce cadre, une analyse de textes peut être utilisée pour examiner les effets d'un texte sur l'extraction des réponses aux questions et l'étendue de leur compréhension. Par ailleurs, représenter un texte arabe sémantiquement peut faciliter ce processus en aidant à comprendre la structure sémantique très compliquée de la langue arabe. Notre approche analyse non seulement un texte donné, mais elle est également efficace pour la représentation logique des documents arabes (questions et textes) en trouvant des structures d'arguments prédictifs de chaque phrase à partir des graphes conceptuels de ces informations.

John F. Sowa en 1984 [Sowa, 1984] a présenté et a défini le formalisme du graphe conceptuel comme un graphe connecté, biparti et fini. En outre, ce formalisme est utilisé comme un langage intermédiaire pour interpréter le formalisme orienté objet et le langage naturel. Le graphe étant constitué par un ensemble de nœuds connectés par des liens; les nœuds du graphe sont soit des concepts (notés par des rectangles), soit des relations conceptuelles (notées par des ovales). Les nœuds de relation conceptuelle indiquent une relation impliquant un ou plusieurs concepts. Les concepts se composent d'un type de concept et d'un référent (instanciation du type de concept); les relations consistent d'un type de relation. Dans le traitement du langage naturel, la modélisation conceptuelle est une façon de modéliser la sémantique. La sémantique des textes se tourne en sémantique des modèles conceptuels à un niveau supérieur d'abstraction, en termes de concepts et relations. Un exemple d'un graphe conceptuel relatif à la phrase «أكمل التلميذ الدرس» (« L'élève a terminé la leçon») est illustré par la figure 3.10.

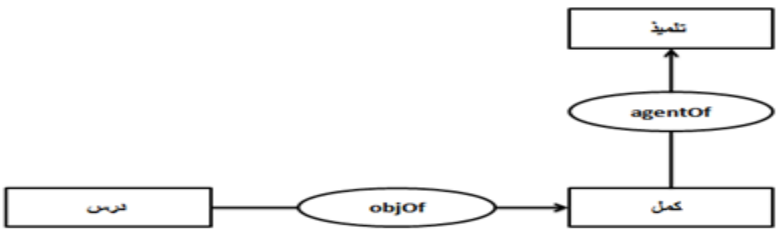


Figure 3.10: Exemple de graphe conceptuel pour « أكمل التلميذ الدرس»

Depuis leur introduction, les recherches reliées au formalisme d'un graphe conceptuel ont connu un essor considérable dans différents domaines (le traitement des langages naturels, les bases de données sémantiques, les systèmes de base de connaissances, les systèmes d'informations, les systèmes multi-agents, l'écriture des spécifications, etc.). D'où, plusieurs approches qui sont focalisées sur ce formalisme sont développées et elles sont actuellement à une étape assez avancée. Ces approches ont conduit à la mise en œuvre de quelques systèmes. Ces systèmes ont utilisé les graphes conceptuels pour la représentation du texte afin de résoudre un problème particulier [Mihalcea & Radev, 2011]. En raison des avantages obtenus en utilisant le formalisme de graphe conceptuel, il a été adopté dans nos travaux pour exprimer la sémantique dans l'analyse de la question et du texte.

Toutefois, quelques systèmes utilisent les graphes conceptuels exclusivement pour la présentation, d'autres au niveau indexation et interrogation. En effet, la représentation de chaque phrase d'un texte est une structure de graphe composé par des concepts reliés entre eux. En fait, le recours à des graphes conceptuels pour représenter des textes ayant une longue histoire dans le TALN, elle s'est concentrée sur les techniques de compréhension de la langue. En particulier, la formulation de textes en termes de graphes conceptuels s'est énormément développée au cours des dernières décennies, particulièrement pour les langues latines telles que l'anglais et le français. En effet, les approches qui se basent sur la représentation sémantique en termes de graphes conceptuels ont déjà été proposées dans le contexte de plusieurs applications de TALN, y compris, la question-réponse.

Comme beaucoup d'autres disciplines, le domaine de la question-réponse traite également des textes qui peuvent être représentés sous forme des graphes conceptuels. En effet, une signification sémantique d'une phrase peut être obtenue en traduisant les graphes conceptuels en des représentations logiques. Cependant, il n'est pas facile de transformer le texte du langage naturel en des structures de graphes conceptuels. D'ailleurs, nous remarquons que certaines langues sont mieux servies que d'autres, en raison de la maturité de la recherche dans les pays qui parlent cette langue. Notamment, nous décrivons quelques recherches qui soulignent la représentation de texte sous formes de graphes conceptuels, en langues latines et nous envisagerons ensuite quelques travaux en relation en arabe. Ces graphes sont générés automatiquement à partir des textes par un traitement linguistique.

Dans un aperçu court mais complet des approches fondées sur les graphes conceptuels pour la question-réponse en anglais, [Mollá et al., 2007] présentent « AnswerFinder », un cadre pour les systèmes de question-réponse qui participe à la tâche (QAst) de CLEF 2007. Mollá et ses associés introduisent leur approche pour apprendre des modèles de graphes qui relient les questions avec leurs réponses convenables. Cette approche est fondée sur la transformation des formes logiques des questions et des réponses candidates en des graphes conceptuels.

[Salloum, 2009] présente un système de question-réponse basé sur le formalisme des graphes conceptuels. Ce système répond aux questions en représentant les connaissances dans les documents et les questions avec le formalisme du graphe conceptuel. Le système proposé traite des questions commençant avec les pronoms wh et «comment». Il trouve les réponses de ces questions en projetant les graphes conceptuels de la question sur les graphes conceptuels des documents et extrait la réponse des résultats de la projection. [Gómez-Adorno et al., 2014] décrivent une approche dans le cadre de leur participation à la tâche de question-réponse 2014 basée sur les examens d'entrée. Cette approche utilise une structure de graphe pour représenter les documents et les hypothèses de réponse. Elle extrait les caractéristiques linguistiques des deux graphes de documents et hypothèses de réponse, en parcourant les chemins les plus courts. Les caractéristiques sont mises en œuvre pour calculer la similitude entre le document et les hypothèses de réponse.

[Jurczyk & Choi, 2015] proposent une approche de représentation de connaissances en des graphes conceptuels pour répondre à des questions complexes. Ces graphes consistent en des relations entre des entités à l'intérieur et à travers des phrases. Dans ce processus, la construction d'un graphe s'effectue en combinant quatre types d'informations: les dépendances syntaxiques, les étiquettes de rôle sémantiques, les entités nommées et les liens de co-références générés par des outils existants. Les vertices d'un graphe représentent des unités linguistiques (par exemple, des mots, des phrases) et des arêtes représentent leurs relations syntaxiques ou sémantiques. [Hixon et al., 2015] décrivent un système de question-réponse qui apprend à partir des dialogues de conversation pour relier les concepts en questions scientifiques aux propositions dans un corpus de faits, il stocke les nouveaux concepts et des relations dans un graphe de connaissance et utilise le graphe pour résoudre les questions. Ce système acquiert des connaissances pour la question-réponse à partir des dialogues ouverts en langage naturel sans une ontologie fixe ou un modèle de domaine

prédéterminant ce que les utilisateurs peuvent dire. [Teney et al., 2016] proposent un réseau neuronal profond pour améliorer la question-réponse visuelle qui traite les représentations structurées en graphe du contenu des scènes et des questions. Cela permet d'exploiter les outils de traitement du langage naturel existants, en particulier les emboutissages de mots pré-formés et l'analyse syntaxique.

S'agissant de la langue arabe, la représentation de textes à l'aide des graphes conceptuels fût déjà l'objet de quelques travaux de recherche. Cependant, ces travaux n'ont pas atteint le même niveau d'avancement que celles concernant les langues latines. Ainsi, la représentation des textes sémantiquement via des graphes conceptuels tenue en compte l'une des techniques actuelles qui facilitent le processus de manipulation de différentes applications de TALN, telles que, la recherche d'informations, la question-réponse, la traduction automatique, le résumé automatique, le raisonnement, etc. A ce titre, Ismail et ses associés proposent une approche pour le résumé de textes arabes [Ismail et al., 2013]. Ces auteurs créent un résumé abstractif pour un seul document d'entrée en langue arabe. Ce résumé est généré par trois modules: convertir le texte arabe d'entrée en un graphe sémantique riche, puis effectuer une réduction du graphe, et finalement générer le résumé de texte à partir du graphe réduit.

[Abouenour et al., 2014] présentent la construction d'une nouvelle ontologie arabe pour les applications de TALN, en combinant l'information lexicale et les relations hyponmiques de WordNet arabe avec les trames sémantiques et syntaxiques des classes verbales dans VerbNet arabe. Cette combinaison permet une représentation sémantique des concepts clés afin d'admettre les systèmes intelligents et le raisonnement sémantique. Les auteurs transforment les trames de WordNet arabe en des graphes conceptuels pour assurer l'utilisabilité de cette ressource dans la question-réponse arabe. L'adoption de la transformation des graphes conceptuels permet de représenter le sens dans les questions et les passages dans le but de comparer leurs représentations. Ainsi, [Bouhriz et al., 2015] proposent une approche qui permet l'extraction de concepts. Ces derniers représentent le contenu sémantique d'un texte arabe. Ces concepts sont extraits à partir de WordNet arabe. Ensuite, les auteurs appliquent à ces concepts une analyse formelle pour produire un ensemble de concepts, plus réduits et plus pertinents. Ces concepts peuvent être utilisés plus tard comme descripteurs sémantiques pour la phase d'indexation.

Finalement, [Nasri et al., 2016] décrivent une approche liée à l'analyse sémantique de la langue arabe. La sémantique est représentée avec le formalisme des graphes conceptuels qui offre un formalisme de représentation de la connaissance puissant et peut supporter de nombreuses applications de TALN et de raisonnement. Cette approche consiste à analyser d'une manière syntaxique le texte arabe avec un analyseur syntaxique, qui marque les constituants du texte avec des rôles thématiques et détermine les motifs syntaxiques des phrases. L'information syntaxique suite à l'information sémantique fournie par l'ontologie construite aide à extraire l'information sémantique d'un texte arabe donné. Enfin, à notre connaissance, notre travail et celui de Abouenour et al., sont les seules tentatives dédiées à la question-réponse arabe qui intègrent une représentation de textes arabes sous formes de graphes conceptuels. Ceci facilite les traitements qui suivent. Enfin, notre méthode proposée pour la construction du graphe conceptuel est décrite dans le chapitre 4.

4.4 **Une interprétation logique pour la compréhension automatique de textes**

La représentation logique a une longue histoire en langage naturel; c'est une étape intermédiaire entre l'analyse syntaxique et sémantique profonde [Moldovan et al., 2002]. Elle a réalisé des progrès considérables dans des domaines clés de traitement du langage naturel; une telle application est la question-réponse où le problème est de trouver des réponses exactes aux questions exprimées en langage naturel en recherchant une grande collection de documents [Voorhees & Harman, 2002]. En revanche, ce type de recherches en intégrant des procédures de raisonnement logique dans certaines applications de traitement de la langue arabe n'a pas encore atteint un stade évolué. Ceci est du, d'une part, à la complexité de cette langue, d'autre part, à l'insuffisance des recherches sur cette langue. Ainsi, la question-réponse avancée nécessite des outils sophistiqués de traitement de texte basé sur les méthodes de TALN et de raisonnement logique [Moldovan et al., 2002].

Dans la littérature de question-réponse, les travaux qui se fondent sur les techniques de théorèmes de démonstration et sur les formes logiques explicites sont, dans une certaine mesure, mis en évidence et/ou utilisés. Ces travaux emploient des formes logiques explicites et des techniques de démonstration de théorèmes. Un certain nombre de ces travaux adoptent des formalismes qui sont fondés sur la logique des prédicats. Ils représentent particulièrement l'hypothèse et le texte sous une forme en logique du premier ordre qui favorise la structure de prédicat-arguments. Cette structure donne une interprétation sémantique des phrases en

déterminant qui a fait quoi, à qui, où, quand, comment et pourquoi. Presque la totalité de ces systèmes sont dédiés pour l'anglais.

Revenons tout d'abord, dans [Harabagiu et al., 2000], les auteurs révèlent un système de question-réponse qui utilise un démonstrateur de théorème basé sur la transformation de forme logique de la question-réponse. Par conséquent, les transformations sémantiques pour les questions et les réponses sont transformées en des représentations logiques et présentées à un démonstrateur de théorème simplifié. En outre, dans [Moldovan & Rus, 2001], les auteurs ont discuté la conversion des gloses WordNet en des axiomes via la transformation de forme logique dans le contexte de WordNet étendu (XWN). Simultanément, dans [Moldovan et al., 2003], les auteurs mettent en œuvre le système COGEX qui prend en question-réponse des formes logiques et des axiomes WXN / TALN et sélectionne des réponses basées sur le score de preuve. En outre, les auteurs, dans [Moldovan et al., 2007], présentent l'optimisation des capacités de COGEX en intégrant l'information sémantique et contextuelle pendant la génération de formes logiques.

De plus, Mollá et ses associés décrivent ExtrAns, un système de question-réponse appliqué au domaine de l'unix. Ce paradigme utilise des formes logiques minimales (MLFs) qui sont converties en des faits / requêtes Prolog [Mollá et al., 2000]. Dans [Mollá, 2003], l'auteur compare les formes logiques minimales avec les relations grammaticales telles que les mesures de score de similarité basées sur le chevauchement pour le classement des réponses. Après cela, Rinaldi et ses collaborateurs ont exploré une approche logique qui est dédiée à la question-réponse biologique en adaptant le système ExtrAns de [Mollá et al., 2000] au domaine de la génomique [Rinaldi et al., 2004]. En assurant cette tâche, les auteurs se sont appuyés sur deux échantillons de documents qui sont liés au domaine : (1) corpus GENIA, et (2) corpus «Biovista» composé de textes intégraux d'articles de revues. Ces textes sont générés à partir de MEDLINE à l'aide de deux listes de termes de semences concernant les gènes et les voies.

En outre, Benamara a développé un système de question-réponse appliqué au domaine du tourisme, appelé WEBCOOP, qui contient des faits, des règles et des contraintes d'intégrité encodés en Prolog, et un ensemble de textes indexés par des formules logiques du premier ordre [Benamara, 2004]. Clark et ses collègues présentent également une approche affichée en des couches pour la représentation de la connaissance contextuelle en logique du

premier ordre, couplée à des mécanismes de raisonnement, pour permettre l'inférence contextuelle et le raisonnement par défaut pour la question-réponse [Clark et al., 2005]. Ainsi, Tari et Baral ont proposé un système de question-réponse qui a utilisé AnsProlog pour la représentation et le raisonnement [Tari & Baral, 2005]. Cependant, Baral et ses collaborateurs présentent un système de question-réponse, qui combine AnsProlog et la programmation logique contrainte pour permettre l'inférence textuelle sur les événements, les actions et les relations temporelles [Baral et al., 2005].

Enfin, [Terol et al., 2007] ont exploré une approche fondée sur la logique, en adaptant un système de question-réponse générique de domaine ouvert au domaine médical. Le traitement des questions-réponses est basé sur la dérivation des formes logiques (FLs) à partir des textes par l'application des techniques de TALN et sur le traitement complexe des FLs dérivées.

Et pourtant, la question-réponse arabe reste un petit territoire exploré. Alors que plusieurs approches ont exploité les connaissances sémantiques dans le processus de génération des réponses, notamment en anglais. D'autres approches ont exploré des représentations logiques et des mécanismes d'inférence. En outre, la logique est le niveau le plus important et difficile du traitement du langage naturel. Egalement, les approches à base de la logique sont un sujet riche de recherche, il y a encore place à l'amélioration. Cette tâche a été appliquée pour de nombreuses langues (anglais, français, etc.). D'ailleurs, l'approche la plus populaire pour la transformation d'un texte en anglais en une représentation logique est basée sur l'identification de la structure syntaxique de la phrase, généralement représentée comme un arbre (« l'arbre d'analyse») qui combine systématiquement les phrases dans lesquelles le texte anglais peut être divisé et dont les feuilles sont associées aux éléments lexicaux. En revanche, ce type d'approches n'est pas encore traité pour l'arabe; cela est du à l'absence des outils nécessaires pour cette langue et ses spécificités.

Une idée similaire a déjà été introduite par [Moldovan & Rus, 2001] pour la langue anglaise. Dans leur travail, ils ont décrit une approche pour transformer les gloses WordNet en des formes logiques. La notation utilisée dans la logique de premier ordre contient des informations syntaxiques comme des arguments de position. La transformation de gloses WordNet en des formes logiques est utile pour la démonstration de théorèmes et d'autres applications. Dans nos travaux de thèse, nous entendons une représentation sémantique

incluant un traitement logique des textes en langage naturel. Notre objectif est double, d'un point de vue mathématique, il s'agit de mieux comprendre la structure des formules logiques du premier ordre, notamment si celles-ci sont utilisées comme étant des représentations sémantiques de textes en langue naturelle. D'un point de vue linguistique, il s'agit de combiner deux modes de représentations utilisés en sémantique et logique des langues naturelles. En effet, notre approche se diffère de celle de Moldovan et Rus par le fait qu'elle traite la langue arabe, s'appuie sur les graphes conceptuels pour la représentation logique et l'utilisation des techniques de RTE pour sélectionner la phrase du texte qui implique la réponse exacte. En effet, pour chaque question, le module de recherche de documents fournit un ensemble de passages composés de plusieurs phrases qui peuvent comporter la réponse désirée. Ainsi, l'entrée de l'étape de la représentation logique est constituée des graphes conceptuels de la question et des phrases qui peuvent la répondre. La sortie de cette étape est la question et les phrases en formes logiques appelées QFL (question en forme logique) et PFL (passage en forme logique).

En effet, l'un des grands défis dans la compréhension du texte tel que la construction d'une représentation cohérente globale du texte, est que beaucoup d'informations nécessaires dans cette représentation est implicite. Ainsi, afin de combler les lacunes et de réaliser une représentation globale, les systèmes de traitement du langage ont besoin d'une grande quantité de connaissance du monde, et la création de ces ressources de connaissances reste un défi fondamental [Clark et al., 2008]. Par ailleurs, le fait de pouvoir représenter un problème est intéressant seulement si nous pouvons utiliser cette représentation afin d'effectuer des tâches intelligentes, comme le raisonnement logique. Ainsi, un autre rôle important de la logique est d'élaborer un mécanisme qui permet à une machine d'effectuer des raisonnements.

Depuis le début des années 60, et à partir du premier essai d'analyse automatique proposé par David Cohen qui est déja l'un des premiers théoriciens du domaine de TALN, des recherches se poursuivent dans le cadre du traitement automatique de la langue arabe. Une particularité de notre approche est de procéder à une représentation logique des textes de la question et de phrases de texte. Nous constituons la représentation logique en une forme de prédicat-arguments qui permet de transformer les représentations sémantiques en des graphes conceptuels (de la question et du passage) en des formes logiques. Le processus de représentation consiste en trois étapes: (i) représentation d'un texte arabe (une question et un passage) via le formalisme des graphes conceptuels, (ii) identification d'un algorithme de

transformation d'un graphe conceptuel en logique de premier ordre, (iii) la représentation logique de la question et la des phrases de textes.

D'ailleurs, la création d'un formalisme du raisonnement logique est l'objet même de la logique, selon le philosophe Thomas Hobbes. Pour le faire, un prédicat est généré pour chaque nom, verbe, adjectif ou adverbe [Moldocan & Rus, 2001]; une expression de prédicat est un graphe de la relation de prédicat-arguments. La mise en œuvre des formes logiques repose sur des informations fournies par les graphes conceptuels que nous construisons. De toute évidence, nous avons mis au point un algorithme qui fournit le processus de transformation et qui crée des prédicats et leur assigne des arguments. Les formes logiques générées à partir des graphes conceptuels sont prises en considération pendant l'étape de détermination d'implications textuelles qui repose sur la logique et la représentation sémantique pour extraire la réponse souhaitée.

Plus précisément, J. Sowa définit un opérateur Φ (Phi) qui transforme chaque élément d'un modèle des graphes conceptuels en un élément de la logique du premier ordre. Cette interprétation donne une sémantique particulière au marqueur générique. La transformation d'un graphe de dépendances en formule de la logique du premier ordre se fait de la façon suivante. À chaque nœud du graphe correspond une variable quantifiée à laquelle est appliqué un prédicat unaire ayant pour nom le mot associé au nœud. Les arcs introduisent une prédication binaire ayant pour nom l'étiquette de l'arc et pour arguments les variables associées à ses nœuds source et destination. La formule finale est la conjonction des prédications ainsi obtenues. Ainsi, pour le graphe de la figure 3.10, nous obtenons la représentation logique suivante :

$$\exists X\ \exists Y\ \exists Z : \textbf{اكمل}(X) \wedge \textbf{التلميذ}(Y) \wedge agentOf(X,Y) \wedge \textbf{الدرس}(Z) \wedge objOf(Z,X)$$

5. La reconnaissance d'implications textuelles

La reconnaissance d'implications textuelles est une tâche permettant de décider, étant donné deux fragments de texte T1 et T2, si la signification de T1 peut être déduite de celle de T2. Cette tâche capture génériquement une large gamme d'inférences qui sont pertinentes pour de multiples applications. Par exemple, un système de question-réponse doit identifier les textes qui impliquent la question. Étant donnée la question " من إخترع الحاسوب الألي ؟ ", le texte " تشارلز بابيج العالم الأول الذي اخترع الحاسوب" implique cette question. De ce fait, la réponse attendue est " تشارلزبابيج ". De la même façon, dans la recherche d'information, les concepts

dénotés par une question doivent être impliqués par les documents réponses pertinents. Dans la récapitulation de documents, une phrase ou une expression superbe, pour être omise du résumé, doit être impliquée par une autre phrase du résumé. Dans l'extraction d'information, l'implication existe entre les différentes variations textuelles qui expriment la même relation cible. Enfin, dans la traduction automatique, une traduction correcte doit être sémantiquement équivalente au texte de base. Ainsi, reconnaître les implications textuelles permet de consolider et de promouvoir la recherche sur le traitement sémantique de la langue naturelle et de poser des bases génériques au développement de ces applications.

5.1 **Motivation**

Un phénomène primordial des langues naturelles est la variabilité dans la verbalisation d'un même contenu sémantique : le même sens peut être exprimé ou impliqué par des textes très différents. De nombreuses applications de TALN comme la question-réponse, la recherche d'information, l'extraction d'information, ou encore le résumé automatique ont besoin de traiter cette variabilité afin de pouvoir reconnaître différentes formulations d'un même sens. Néanmoins, il est difficile de comparer, sous la forme d'une évaluation générique, les méthodes sémantiques qui ont été développées dans des applications différentes. Le challenge RTE vise à pallier ce problème en déterminant un cadre d'évaluation permettant de mesurer les capacités sémantiques d'un système sur la base d'une tâche précise, à savoir, la reconnaissance d'implication textuelle.

Plus généralement, le traitement de l'implication textuelle permet de déterminer des relations d'implication, d'équivalence et de contradiction entre des fragments de textes aussi de raisonner sur la signification de ces textes. Au cours des trois dernières années, l'évolution de la campagne RTE (Recognising Textual Entailment) a clairement montré que la reconnaissance d'implications textuelles passe par un traitement profond des données textuelles [Dagan et al., 2005]. L'analyse de l'implication textuelle a montré que si la phrase du texte implique la question, il est nécessaire que les informations contenues dans la question soient présentes dans la phrase ou puissent en être déduites. Ainsi, pour déterminer si une phrase du texte implique une question, nous transformons les représentations sémantiques normalisées en formules logiques puis nous testons la validité de l'implication logique en logique du premier ordre.

D'ailleurs, l'implication textuelle est liée à l'implication logique et l'inférence sémantique [Dagan et al., 2005]. Elle semble que l'effort vise à reconnaître ce que signifie implication aux niveaux lexicaux et syntaxiques, plutôt que d'aborder les questions logiques relativement délicates. Par conséquent, l'absence de cet effort, en particulier dans la langue arabe, et la croissance de la technologie de reconnaissance d'implications textuelles ces dernières années nous ont dirigés à proposer une nouvelle approche pour l'arabe. Notre approche proposée est basée principalement sur la logique, l'implication textuelle et la sémantique pour avoir la possibilité d'identifier une réponse précise à une question dans un langage naturel. Particulièrement, il s'agit ainsi de pouvoir effectuer des raisonnements sur plusieurs passages de textes, et d'intégrer des mécanismes de compréhension de textes, comme la capacité à effectuer des inférences, dans les systèmes de question-réponse.

5.2 **Applications de l'implication textuelle**

L'implication textuelle est un domaine de recherche moderne en traitement du langage. Il a pour but de fédérer les recherches en TALN afin de proposer des méthodes de traitement du langage au niveau lexical, syntaxique et sémantique indépendamment dans un large éventail d'applications de traitement du langage naturel, y compris, la question-réponse, le résumé automatique, la génération de texte, la traduction automatique, la recherche d'informations, etc., [Dagan et al., 2013]. Cette tâche a été introduite par Dagan et Glickman dans la première campagne d'évaluation appelée Recognizing Textual Entailment (RTE). L'objectif de cette tâche est de modéliser la capacité humaine à savoir si une hypothèse peut être déduite à partir d'un texte. Plus précisément, elle vise à déterminer automatiquement si un segment de texte (H) est déduit d'un autre segment de texte (T) [Dagan et al, 2005]. Depuis 2005, RTE a été proposée comme une tâche dont le but est de capturer les principaux besoins d'inférence sémantique entre les applications en linguistique computationnelle [Dagan et al., 2009]. Une nouvelle campagne d'évaluation a lieu chaque année avec différentes redéfinitions de la tâche. La RTE est utile aux systèmes de recherche d'informations dans le sens où un document est pertinent s'il implique la phrase utilisée comme une requête. Pour les systèmes de question-réponse, un texte est une réponse à une question si la clôture existentielle de la représentation de la question est impliquée par la représentation de ce texte. Pour le résumé automatique, elle permet de vérifier que l'hypothèse est bien impliquée par le texte et d'éviter les redondances entre phrases (une

phrase impliquée par une autre phrase présente dans le résumé ne sera pas incluse) [Dagan et al., 2006].

5.3 **Implication textuelle en question-réponse**

Comme était mentionné précédemment que la problématique d'implication textuelle est applicable dans les systèmes de question-réponse mais se retrouve également dans d'autres tâches de traitement automatique des langues naturelles telles que la paraphrase, le résumé de texte, etc. En question-réponse, cette technique est fréquemment utilisée pour valider une réponse récupérée par un système de question-réponse. Que se soit dans le défi RTE du PASCAL ou dans la tâche AVE du CLEF [Peñas et al., 2010], le problème de la question-réponse est modélisé en considérant une question Q transformée en une phrase affirmative en tant qu'hypothèse et un passage de texte contenant une réponse candidate A en tant que texte (e.x. les systèmes déterminent alors si A implique Q). Par exemple, un système de question-réponse doit identifier les textes qui impliquent une réponse hypothétique. Compte tenu de la question suivante: "من الذي رسم "الصرخة؟", le texte: " لوحة الصرخة الشهيرة لإدفارد مونش التي رسمها عام 1893", implique la réponse hypothétique suivante :"إدوارد مونش هوالذي رسم لوحة الصرخة".

En outre, il existe deux grands types d'approches qui ont été préalablement conçues pour la détection de l'implication textuelle. D'une part, les approches analytiques s'appuient sur un formalisme de représentation et des analyses profondes syntaxiques ou sémantiques. En effet, ces approches recherchent une similarité globale entre la forme syntaxique de la question et celle du passage en considérant par exemple les transformations à effectuer pour passer d'une forme à une autre. Ce type de formalisme permet aussi de définir des vérifications, telles que la recherche de liens syntaxiques communs au passage et à la question. Les approches qui s'intéressent davantage à l'analyse sémantique de la question et des passages et, à partir de celle-ci, créent un formalisme logique à partir duquel un système de preuves peut être appliqué afin de détecter l'implication. D'autre part, les approches fondées sur la combinaison par apprentissage de différents critères locaux d'ordre lexical et syntaxique.

En effet, plusieurs expériences ont été menées pour détecter l'implication textuelle pour de nombreuses applications du TALN, notamment pour la question-réponse. La première d'entre elles, présentée dans [Glickman et al., 2006], part du principe que, pour que le texte implique l'hypothèse, il faut qu'il contienne les mots de l'hypothèse à l'identique ou sous

forme de variante. L'étude se penche ainsi sur les implications lexicales et quatre phénomènes ont été révélés ainsi que leurs fréquences d'occurrences. Le système COGEX [Tatu et al. 2006] traite la validation de réponses en effectuant une implication sur le sens. Le système se fonde sur l'idée que le texte implique l'hypothèse s'il implique logiquement son sens. Dans ce formalisme les prédicats correspondent aux verbes, noms et adjectifs. Les relations sont obtenues également par une analyse syntaxique. Un ensemble d'axiomes venant de la ressource entendue de WordNet [Mihalcea & Moldovan 2001] est également considéré.

Ainsi, l'approche de [Fowler et al., 2005] utilise le système COGEX (une version modifiée du prouveur OTTER) pour reconnaître les implications textuelles. Les auteurs utilisent ensuite des axiomes qui permettent d'exprimer les équivalences syntaxiques et d'affaiblir la complexité des autres formules logiques. D'autres types de systèmes s'appliquent sur l'anglais ont recours à WordNet [Fellbaum, 1998] pour détecter le rapprochement entre deux mots. La méthode la plus simple, présentée notamment dans [Pakray et al. 2009] ou [Ofoghi, 2009], est basée sur les relations de synonymie ou d'hyperonymie. Plusieurs systèmes tiennent compte des entités nommées pour détecter la validité des réponses ou l'implication textuelle. [Ferrández et al., 2009] définissent ainsi deux critères : le premier teste si toutes les entités nommées de l'hypothèse sont trouvées dans le texte, le second correspond à la proportion d'entités nommées de l'hypothèse présentes dans le texte.

Le système de l'université d'Hagen ([Glöckner, 2007], [Glöckner & Pelzer, 2008]) contient aussi un mécanisme par preuve logique mais débute par une étape visant à normaliser le texte en modifiant les mots afin qu'ils soient au plus près de ceux de l'hypothèse. La preuve est effectuée par un mécanisme de relaxation récursive. Tout d'abord, le système prend l'ensemble des prédicats du texte et de l'hypothèse et détermine s'il y a une implication. Si ce n'est pas le cas, il retire des prédicats de l'hypothèse. Ce mécanisme est exécuté jusqu'à obtenir une implication. La décision finale est alors faite à partir des mots présents dans cette nouvelle hypothèse.

En comparaison avec l'anglais, la langue arabe a relativement moins d'attention pour la reconnaissance de l'implication textuelle, en raison des défis auxquels nous pouvons faire face en impliquant un texte à partir d'un autre. L'un de ces défis est que la langue arabe a une morphologie dérivée productive, où à partir d'une seule racine, plusieurs formes pourraient

être dérivées. Ces mots dérivés deviennent confus au cas où des diacritiques manquent [Alabbas, 2011]. L'arabe est l'une des langues les plus difficiles à traiter en raison de sa richesse morphologique et de son ordre de mots relativement gratuit ainsi que de sa nature diglossique (où la norme et les dialectes se confondent dans la plupart des genres de données) [Almarwani & Diab, 2017]. En outre, cette langue nécessite encore de ressources informatiques artisanales à grande échelle qui ont été très utiles pour l'anglais, comme les ontologies. Donc, il y a une bonne portée d'amélioration.

D'ailleurs, il existe peu de recherches publiées dans la littérature de reconnaissance de l'implication textuelle arabe, qui a motivé le travail sur ce thème. A notre connaissance, le travail de [Alabbas, 2011] a été le premier à cibler cette question dont son objectif est de mettre en évidence le système ArbTE (Arabic Textual Entailment) pour évaluer les techniques d'implication textuelle existantes lorsqu'elles sont appliquées à l'implication textuelle arabe. En effet, la technique adoptée par l'auteur correspond à des paires d'hypothèses de texte utilisant l'algorithme de distance d'édition d'arbre (TED). Les auteurs dans [Alabbas & Ramsay, 2013] ont proposé l'utilisation de la distance d'édition d'arbre étendue avec des sous-arbres, ce qui donne un algorithme d'appariement plus flexible pour identifier l'implication textuelle en arabe.

De plus, [AL-Khawaldeh, 2015] a examiné la négation et la polarité en tant que caractéristiques supplémentaires pour la reconnaissance d'implication textuelle arabe. Il s'est concentré sur l'importance de la classification des contradictions dans les systèmes de RTE, car la contradiction pourrait inverser la polarité actuelle de la phrase. Les mots de contradiction sont traités comme des mots vides et éliminés du texte dans le stade du prétraitement avant que le stade de l'implication soit terminé. [Khader et al., 2016] ont proposé une méthode d'implication textuelle qui a été construite en utilisant le langage Python. Cette méthode est basée sur une combinaison lexicale et sémantique. Elle comprend deux phases (Calcul de chevauchement de mots et vérification de Bigram Matching). Récemment, [Almarwani & Diab, 2017] ont posé le problème de l'implication textuelle comme une tâche de classification binaire. Sans dépendre des ressources externes, les auteurs utilisent à la fois des caractéristiques traditionnelles et des représentations distributives pour identifier si un T implique une H.

Dans notre cas, l'implication textuelle est considérée comme étant un problème d'implication logique entre les sens des deux phrases [Tatu et al., 2006]. A ce niveau, l'implication entre deux phrases est acceptée si leur sens se concorde. En d'autres termes, le prédicat peut posséder divers types d'arguments (p.e.x. sujet, complément d'objet direct et complément d'objet indirect). Pour cela, la structure prédicat-arguments est souvent utilisée, c'est-à-dire que, les phrases de textes T et les questions considérées comme hypothèses H sont transformés en un ensemble de prédicat et leurs arguments à travers un algorithme de transformation à fin de déduire l'implication après avoir être transformées en des graphes conceptuels. Par conséquent, pour déterminer si un texte T1 implique textuellement un texte T2, nous traduisons les représentations sémantiques présentées par des graphes conceptuels en des formules logiques. Puis, nous testons la vérification de l'implication logique. Étant donnés deux textes T1 et T2, T1 implique T2 si et seulement si (la traduction en logique du premier ordre de) l'une des représentations sémantiques normalisées de T1 implique (la traduction en logique du premier ordre de) l'une des représentations sémantiques normalisées de T.

5.4 **Traitement logique de l'implication textuelle**

Le principe du traitement logique est la relation de conséquence logique entre les textes, l'idée qu'un texte découle logiquement d'un autre texte. Ainsi, les formes logiques sont les premières représentations logiques de commande de texte en langage naturel, la notation est très proche de la langue naturelle. Une forme logique est une collection d'instances de prédicats dérivés du texte. Dans nos travaux de thèse, nous proposons une approche pour transformer une phrase présentée selon le formalisme d'un graphe conceptuel en une forme logique. Ceci est effectué via un algorithme de transformation décrivant le principe de l'opérateur Φ (Phi) de Sowa. Ce dernier propose d'associer à chaque graphe conceptuel G une formule bien formée Phi (G) du calcul de prédicat de premier ordre. Par conséquent, la transformation en des formes logiques est utilisée pour reconnaître l'implication textuelle entre une question et les phrases du texte qui lui répond. Par conséquent, notre travail exige la compréhension automatique de textes arabes à un niveau plus profond. Donc, le principal enjeu est de savoir comment convertir une information trouvée dans un texte donné en une langue avec laquelle une machine peut raisonner et faire des décisions, et facilement suivre ces questions en langage naturel d'un utilisateur. En effet, notre objectif est de permettre une analyse approfondie de textes qui puisse aller jusqu'à une représentation sémantique et logique, sans que ce soit systématiquement une obligation. Pour

le faire, nous adoptions une représentation sous forme de graphes conceptuels et de prédicat-arguments.

À notre connaissance, il existe quelques approches logiques pour la tâche de la RTE. En outre, bien que cette tâche a été définie beaucoup moins rigoureuse que l'implication logique, Harabagiu et ses associés croient que la RTE entre une question et un ensemble de réponses candidates peut permettre aux systèmes de question-réponse d'identifier les réponses correctes avec une plus grande précision qu'avec les mots clés ou des techniques basées sur le modèle [Harabagiu & Hickl, 2006]. Nous continuons à poursuivre une approche sémantique et logique qui permet de convertir la question et les phrases de texte répondant à cette question en une représentation sémantique via les graphes conceptuels en des représentations logiques puis de chercher si une de ces représentations de phrases d'un passage implique la représentation logique de la question en utilisant une technique de RTE. Le système doit identifier le segment de texte qui contient la réponse. L'implication entre chaque phrase du texte T et la phrase H de la question peut aider à détecter le segment qui contient la réponse.

Conclusion

Au cours de ce chapitre, nous avons présenté, en premier lieu, le terrain pour une nouvelle approche dédiée à la question-réponse arabe. Cette approche se fonde sur la combinaison des techniques de TALN, RI, raisonnement automatique, RTE, etc. Nous nous sommes focalisés particulièrement sur deux méthodes proposées pour l'analyse des questions et la construction de notre corpus de questions-textes. En deuxième lieu, nous avons présenté les fondements théoriques pour une nouvelle approche logique pour la question-réponse en arabe, à savoir, la compréhension automatique d'un texte donné et la reconnaissance d'implications textuelles. Nous avons défini les raisons motivant les recherches dans ce domaine. Au niveau de l'analyse des questions, avons présenté les caractéristiques des questions qui sont extraites par le module d'analyse des questions. Ces caractéristiques sont utilisées pour rechercher des documents à partir du Web et pour construire notre corpus.

Dans le chapitre suivant, nous allons détailler les étapes de l'approche proposée qui consiste à trouver une réponse précises à une question en langue arabe. Cette approche se fonde absolument sur une représentation sémantique et logique. Elle sera implémentée et évaluée sur le corpus construit de questions-textes collectés à partir du Web.

CHAPITRE 4: UNE NOUVELLE APPROCHE SEMANTIQUE ET LOGIQUE POUR LA QUESTION-REPONSE ARABE

Introdution……..... 104

1. Motivations 104

2. Démarches de l'approche proposée 105

2.1 Analyse des questions 106

2.1.1 Prétraitement des questions 107

2.1.2 Transformation des questions 107

2.1.3 Traitement linguistique des questions 108

2.2 Recherche des documents 110

2.2.1 Identification des documents pertinents 110

2.2.2 Sélection des passages pertinents 111

2.3 Analyse des passages 112

2.3.1 Nettoyage des passages 112

2.3.2 Normalisation des passages 112

2.3.3 Segmentation des passages 113

2.3.4 Traitement linguistique des passages 114

2.4 Représentation logique 115

2.4.1 Construction d'une représentation sémantique avec les graphes conceptuels 116

2.4.2 Raisonnement logique à l'aide des graphes conceptuels 127

2.4.3 Transformation des graphes conceptuels en des représentations logiques 129

2.4.4 Détermination de l'implication textuelle 131

2.5 Recherche de la réponse précise 140

2.5.1 Extraction et pondération des réponses candidates 140

2.5.2 Sélection de la réponse précise 141

Conclusion……..... 142

Introduction

Une nouvelle ligne de recherche en question-réponse arabe est l'intégration de la sémantique, la logique et l'implication textuelle. Dans cette thèse, nous portons notre attention sur la proposition d'une nouvelle approche qui est fondée sur la reconnaissance d'implication logique. Notre approche sémantique et logique sert à répondre à des questions qui combine l'utilisation des techniques de recherche d'informations, de traitement automatique de la langue naturelle et d'intelligence artificielle destinée à un système de question-réponse arabe. L'utilisation de cette approche nous permet d'envisager la compréhension automatique des textes arabes. Une telle sorte de compréhension fournit une représentation sémantique et logique des questions et des passages réponses. L'idée est de représenter chaque question et son passage réponse avec le formalisme des graphes conceptuels. Puis, transformer ces graphes en des représentations logiques via un algorithme de transformation. Enfin, détecter les implications textuelles entres ces représentations logiques.

1. **Motivations**

Notre approche se fonde essentiellement la compréhension automatique de textes pour trouver la réponse correcte à une question donnée. Pour le faire, nous proposons d'abord de transformer ces textes en des graphes conceptuels puis de convertir ces graphes en des formes logiques. Enfin, nous proposons de déterminer les implications textuelles entre ces représentations logiques.

Par conséquent, notre approche est motivée par les observations suivantes :

- La langue arabe est moins préoccupée par les chercheurs dans le domaine de la question-réponse [Abufardeh & Magel, 2008], [Al-daimi & Abdel-amir, 1994], [Habash & Rambow, 2005], [Akour et al., 2011].
- La quasi-totalité des systèmes de question-réponse arabe se sont concentrés sur des approches morphosyntaxiques et le raisonnement sémantique et /ou logique a un très peu d'intérêt pour les approches proposées.
- A notre connaissance, aucun système de question-réponse qui couvre une approche logique en arabe [Abouenour et al., 2012].
- La représentation des questions et des passages réponses au moyen du formalisme des graphes conceptuels [Sowa, 1984] facilite la représentation logique de ces informations.

- Nouvelles tendances de recherches pour intégrer en arabe la logique, l'inférence, la compréhension en lecture, etc., exemple: l'étude de [NBdour & Gharaibeh, 2013].
- Les systèmes de question-réponse, surtout ceux développés en arabe, se sont focalisés sur la syntaxe et la morphologie pour analyser les questions et les documents. Cependant, à notre connaissance, une seule tentative qui se fonde sur la logique pour la représentation des questions en des formes logiques.
- Il reste encore beaucoup à faire dans le domaine de reconnaissance d'implication textuelle (RTE), l'inférence et la logique. En effet, la technique de reconnaissance d'implication textuelle (RTE) n'est pas bien étudiée dans la question-réponse arabe comme les autres langues, notamment l'anglais.

2. Démarches de l'approche proposée

Les sous-sections présentent une description détaillée de notre approche [Bakari et al., 2014]. Cette approche consiste en six principales étapes ; la première traite l'analyse de la question, la deuxième prend en considération la recherche de documents, la troisième aborde l'analyse de documents retournés, la quatrième donne un formalisme de représentation particulier de la question et du passage, à savoir la représentation sémantique et logique. La dernière s'attaque à la sélection de la réponse précise qui est considérée comme un problème d'implication textuelle.

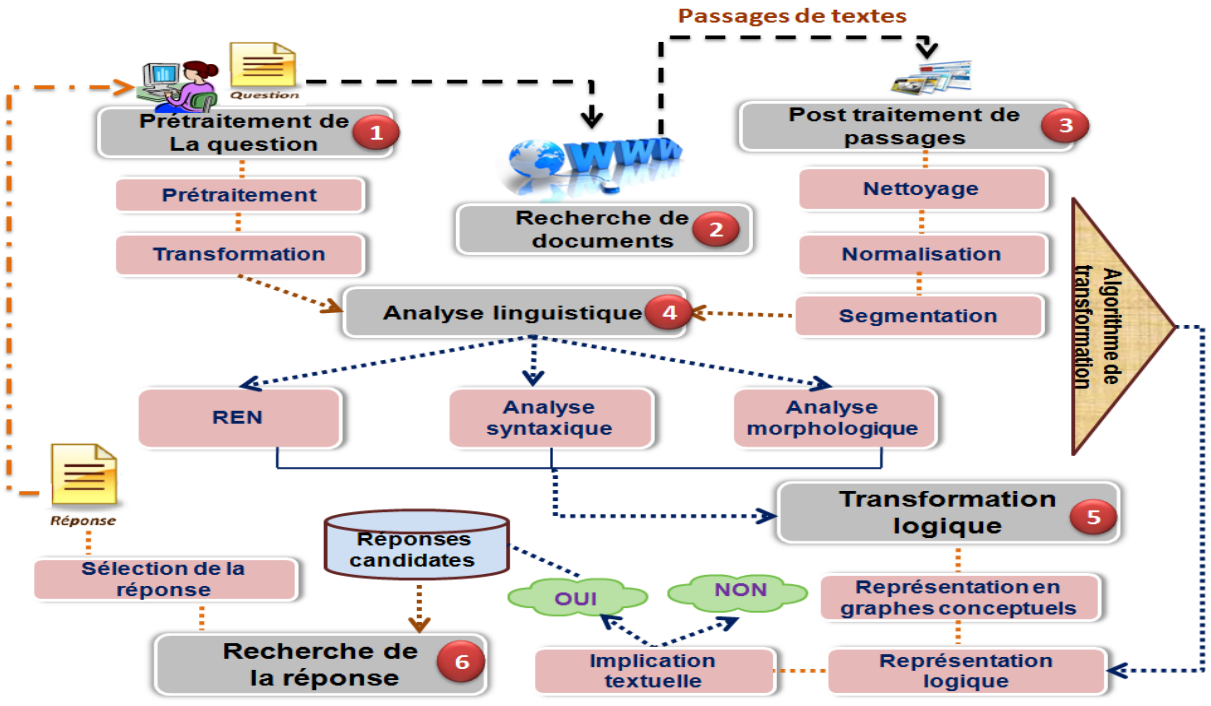


Figure 4.11: Les étapes de l'approche proposée

2.1 **Analyse des questions**

Dans cette section, nous présentons en détails l'étape d'analyse de la question tout en présentant leurs diverses caractéristiques qui peuvent nous aider à trouver les passages réponses pertinents et à sélectionner la réponse précise. En effet, bien que les techniques diffèrent d'un système à l'autre, la plupart des systèmes de question-réponse est basée sur une étape d'analyse de la question. Cette étape est définie comme étant une étape préliminaire dans le processus de recherche des réponses précises à des questions en langue naturelle. Dans notre contexte, cette phase est une succession de trois étapes, le résultat de chaque étape sera exploité par la suivante. En général, les caractéristiques extraites de cette étape facilitent l'extraction de la réponse précise. Toutes les informations obtenues par ce module sont données aux étapes suivantes du système [Rodrigo et al., 2010]. En arabe, la majorité des études se concentre sur l'extraction de mots-clés et la reconnaissance des entités nommées. Dans notre cas, nous ajoutons la reformulation de la question sous une forme déclarative. Celle-ci est utilisée prochainement pour construire des graphes conceptuels et générer des formes logiques.

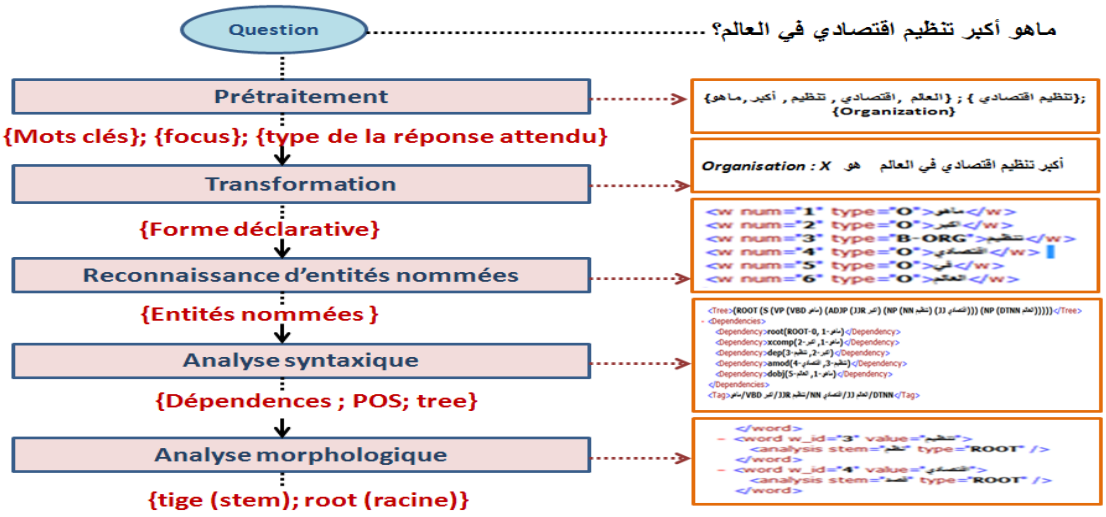


Figure 4.12: Les étapes d'analyse de la question

Afin de répondre à une question en langage naturel, plusieurs caractéristiques des questions ont été plus au moins mises en évidence et/ou utilisées dans notre étude sur des questions réelles. Nous nous concentrons sur un type particulier de question, à savoir les questions factuelles. Notre module d'analyse de la question suppose que chaque question doit être une phrase déclarative simple qui est composée d'une séquence de mots et cherche le type de réponse attendue comme une preuve utile pour extraire la réponse précise. Nous décrivons le processus d'analyse avec des exemples de 5 types de questions collectées dans

notre corpus. Étant donné un exemple concret d'analyse de la question et des traitements mis en jeu. Considérons la question suivante : « **ماهو أكبر تنظيم اقتصادي في العالم؟** », son traitement par l'analyse de la question est décrit dans [Bakari et al., 2017] ; la figure 4.12 présente la sortie de chaque étape.

2.1.1 Prétraitement des questions

La première étape en vue d'analyser la question subit des tâches de prétraitement. Cette étape tente à déterminer des caractéristiques principales de la question (e.g. les mots-clés aussi le type de la réponse attendue) pour toute question factuelle. Ces caractéristiques peuvent nous aider plus tard à récupérer les passages de textes pertinents à partir du Web. Elles sont également utilisées par d'autres modules pour générer la réponse exacte (sélection de la réponse précise). Même dans les autres études de question-réponse arabe, dans la phase d'analyse des questions, des tâches de prétraitement ont été appliquées afin d'éliminer les données non pertinentes (les mots vides sont éliminés et les particules d'interrogation sont supprimées). Plus précisément, le prétraitement produit ces principales caractéristiques avant la transformation de la question:

- ***Les mots clés* :** Il est nécessaire d'analyser les questions au delà du découpage en mots-clefs qui sont obtenus en éliminant tous les mots de vides par comparaison de chaque mot reconnu avec les éléments une la liste des mots vides.
- ***Le type de la réponse attendue :*** Ce type correspond à l'entité nommée attendue en réponse. Selon les réponses, il y a cinq catégories de questions telles que: lieu, personne, date, organisation et expression numérique. Par exemple, si le pronom interrogatif est « Qui » la question attend comme réponse une « personne ».

2.1.2 Transformation des questions

L'analyse de la question est principalement axée sur une éventuelle réécriture. En effet, pour toute question, nous avons généré une reformulation qui est une représentation en sa forme déclarative, voir figure 4.13. Cette transformation pourrait être utile dans le module de représentation logique. En particulier, nous avons éliminé les caractères spéciaux et les particules d'interrogation juste pour obtenir le contenu textuel des questions. Les informations éliminées sont considérées comme des informations non importantes. Elles sont retirées afin d'obtenir des résultats plus significatifs.

Figure 4.13: Représentation de la question en une forme déclarative

2.1.3 Traitement linguistique des questions

Cette étape est divisée en trois sous-étapes: la reconnaissance des entités nommées, l'analyse syntaxique et l'analyse morphologique.

- **La reconnaissance des entités nommées**

La reconnaissance des entités nommées est une des technologies de traitement automatique de langues naturelles les plus prises en compte [Nanda, 2014]. Plusieurs études ont été menées pour montrer l'importance prédominante de REN pour d'autres tâches de TALN, telles que la traduction automatique [Babych & Hartley, 2003], le résumé automatique [Kabadjov et al., 2013], la question-réponse [Mollá et al., 2006], etc. C'est une sous tâche de la piste d'extraction d'informations qui aide à catégoriser et à produire quelques expressions linguistiques mono-référentielles et autonomes [Ehrmann, 2008]. Ces expressions rassemblent traditionnellement à l'ensemble des noms propres (noms de personnes, de lieu et d'organisation) et à certaines expressions numériques et temporelles (expressions de dates, de temps, de pourcentages, etc.).

Les systèmes de question-réponse pourraient bénéficier considérablement de la REN, parce que la réponse à de nombreuses questions factuelles impliquent des entités nommées [Trigui et al. 2012] (par exemple, des réponses à des question (Who) portent fréquemment sur des personnes ou des organisations, les questions (Where) impliquent des locations, et les questions (When) concernent des expressions temporelles) [Brini et al., 2009]. Pour leur analyse, la REN peut être utilisée de manière à reconnaître les éléments au sein de la question qui nous aideront plus tard à trouver les documents pertinents et la construction de passages pertinents comme réponse [Mollá et al., 2006], [Osama et al., 2011], [Abouenour et al., 2012]. Les entités nommées apparaissant dans la question peuvent jouer un rôle important dans l'extraction des réponses possibles.

Dans nos travaux, nous utilisons un outil de reconnaissance des entités nommées ArNER. Nous avons choisi ce système car c'est le standard le plus connu parmi quelques systèmes d'analyse réalisés dans notre laboratoire. Par exemple, la figure 4.14 montre la structure d'annotation de texte nettoyé, elle indique ainsi l'extraction des entités nommées trouvées dans notre exemple de question **Q1:** « **ماهو أكبر تنظيم اقتصادي في العالم؟** » **(**«What is the largest economic organization in the world? »). Plus précisément, cette étape reçoit le texte de la question segmenté qui est déjà généré par la première étape et fournit un fichier XML qui contient toutes les entités nommées.

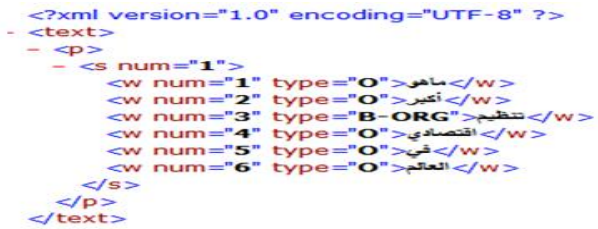


Figure 4.14: Liste des entités nommées extraits par ArNER pour la question Q1

L'analyse syntaxique

Avant de travailler sur l'analyse morphologique d'une question, nous avons choisi d'en étudier sa syntaxe. Certains outils qui garantissent cette analyse syntaxique du texte arabe existent déjà, parmi lesquels l'analyseur Stanford[15] que nous choisissons grâce à sa disponibilité et à la disponibilité de sa documentation. L'étape d'analyse commence par identifier les constituants du texte avec des rôles thématiques et fournir les motifs syntaxiques des phrases. En effet, Les tag des mots aident à minimiser les synsets à partir de Wordnet arabe par l'extraction seulement des synsets ayant le même tag que le mot. Ainsi, les dépendances et les tags nous aident à trouver les relations entre les mots et donc de construire le graphe conceptuel. Les informations obtenues par stanford pour l'exemple de la question **Q1** sont mentionnées dans la figure 4.15.

Figure 4.15: Sorties de Stanford pour la question Q1

[15] https://nlp.stanford.edu/software/lex-parser.shtml

- **L'analyse morphologique**

Dans ce cadre, l'analyse morphologique des mots de la question d'entrée est effectuée en utilisant Khoja Stemmer [Larkey & Connell, 2001]. En fait, Khoja Stemmer supprime le suffixe et le préfixe les plus longs. Il correspond alors au mot restant avec des modèles verbaux et nominaux pour extraire la racine. Il utilise plusieurs fichiers de données linguistiques tels qu'une liste de tous les caractères diacritiques, des caractères de ponctuation, des articles définis et 168 mots vides. Une implémentation Java de l'algorithme de Shereen Khoja est accessible sur le Web[16]. Par exemple, les mots de la question **Q1** et leurs tiges ou racines, ou les deux, sont enregistrés dans des fichiers XML, comme le montre la figure 4.16.

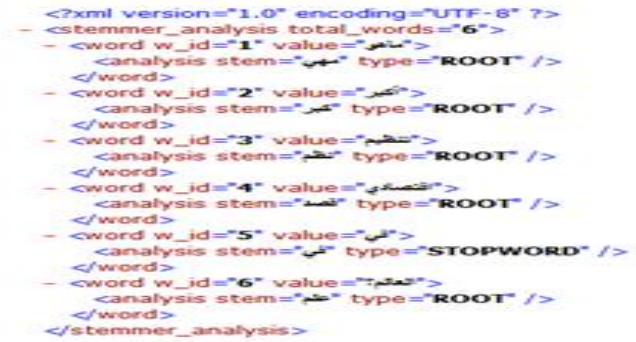


Figure 4.16: Analyse morphologique de la question Q1 avec Khoja Stemmer

2.2 Recherche des documents

Si une analyse de la question est préalablement effectuée, la recherche de documents est effectuée en utilisant le Web. En effet, le Web est devenu le principal référentiel d'informations: presque toutes sortes d'informations (bibliothèques numériques, journaux collections, etc.) dans plus de 1500 langues sont disponibles sur le Web en un format électronique [Rosso et al., 2005]. Tandis que les moteurs de recherche permettent de récupérer des documents sur un thème général, les systèmes de questions-réponses sont employés pour récupérer une information précise, qui tient en quelques mots [El Ayari, 2007].

2.2.1 Identification des documents pertinents

L'identification des documents pertinents consiste à trouver des fragments de documents, susceptibles de contenir la réponse à une question. L'idée retenue consiste à générer de façon automatique les passages qui contiennent un mot particulier ou de ne choisir que les passages contenant des variations des mots de la question. En outre, cela a demandé

[16] https://sourceforge.net/projects/arabicstemmer/

un système de question-réponse pour une réponse courte, propre et bien définie [Pudaruth, et al., 2016]. Pour une question donnée, la plupart des systèmes de question-réponse extraient un nombre important de documents susceptibles de contenir la réponse. Ainsi, ils considèrent que les passages retenus représentent un bon compromis entre un ensemble de documents et des réponses exactes. Par conséquent, établir une réponse précise à une question donnée dans un grand ensemble de documents (au lieu de recherche par mots-clés) réponses (à la place de documents) [Chavan & Gore, 2016] est un enjeu majeur.

2.2.2 Sélection des passages pertinents

Les mots clés de la question détectés lors de son analyse sont fournis au moteur de recherche Google afin d'obtenir les passages de textes pertinents. Usuellement, avec les moteurs de recherche tels que Google, Yahoo et autres, la recherche d'informations est basée essentiellement sur des mots clés pour trouver des documents. Cependant, avec l'énorme quantité d'informations disponibles via les pages Web, ce que l'utilisateur a vraiment besoin est une réponse à sa demande au lieu de documents ou des liens vers ces documents. La méthode pour la recherche de passages de textes pertinents pour chaque question est semblable à celle-ci décrite dans notre travail récemment publié [Bakari et al., 2016]. En particulier, nous interrogeons Google pour réaliser la récupération des documents ou des passages qui pourraient répondre à la liste des questions. En effet, des passages de texte sont un intermédiaire important entre documents complets et des réponses exactes [Brini et al., 2009]. De ce fait, l'existence des mots de la question dans les passages générés peut pointer la présence de la réponse à cette question. Certains systèmes de question-réponse utilisent des moteurs de recherche spécifiques comme Indri [Metzler & Croft, 2004] ou Lucene.

Pour constituer la collection de textes, nous avons récupéré 250 textes renvoyés par le moteur de recherche Google pour les 250 questions analysées précédemment. Le résultat de cette étape se présente généralement sous forme de textes comprenant de plusieurs segments. Chaque segment est un passage de textes qui contient potentiellement la réponse. En effet, un passage est définit comme étant une séquence de longueur fixe de mots commençant et se terminant quelque part dans un document retrouvé [Ofoghi et al., 2006]. Considérons la question suivante: «**متى ولد حنبعل؟**», un passage du texte retenu devrait contenir un passage réponse **P1** comme **: « حنّبعل هو من أعظم القادة العسكريين الذين عرفهم التاريخ ولد بقرطاج سنة 247 قبل الميلاد»**. Une fois les passages intéressants sont trouvés, nous nous sommes également intéressés à les analyser.

2.3 **Analyse des passages**

Dans cette section, diverses étapes sont étudiées pour analyser les passages. D'abord, un post traitement est achevé pour nettoyer le texte source qui est déjà au format html et extrait à partir du Web pour produire un texte en arabe en format txt. Puis, une normalisation est effectuée afin de préparer le texte pour l'étape d'analyse. Ensuite, une segmentation établit le découpage de texte en des unités lexicales (phrases et mots). Enfin, des étapes d'analyse linguistique, à savoir, la REN, l'analyse syntaxique et morphologique sont mises en considération.

2.3.1 Nettoyage des passages

Usuellement, l'utilisateur pose une question en langage naturel sans connaître la structure des sources à interroger. Les textes extraits du Web ne sont pas structurés et exprimés en langage naturel qui est extrêmement difficile à modéliser, ils peuvent également contenir des fautes d'orthographe, de grammaire, ou encore être rédigés dans un style inconnu à l'avance [Belguith et al., 2007]. Il est ainsi possible de leur appliquer, préalablement à leur utilisation, un ensemble de traitements visant à rendre les processus suivants plus rapides ou à normaliser les documents. Ceci entre dans le cadre d'analyser un texte pour découvrir l'information « essentielle » contenue dans ce texte et permet de répondre à une question donnée. En effet, nous constatons que les textes générés à partir du Web contiennent généralement beaucoup de bruits et composants non informatifs, à savoir, les balises HTML, les scripts, etc. De plus, il y a beaucoup d'informations dans le texte qui n'ont pas un sens (les translittérations). Au-delà de ce premier point, et pour établir une réponse précise à une question, l'étape d'analyse des textes recueillis du Web devrait essentiellement débuter par une phase de nettoyage. Cette tâche facilite la transformation de données textuelles en une forme appropriée qui permet un traitement ultérieur et la préparation du texte pour des traitements prochains. Elle prend essentiellement un document html et produit un texte arabe en un format txt.

2.3.2 Normalisation des passages

Les passages générés du web peuvent contenir des mots étrangers, des caractères spéciaux, des nombres (e.g. ‘,:,?, \, $, etc). Ainsi, dans ces textes, certains mots sont très communs et n'ont pas de sens supplémentaire pour leurs contenus réels. Ces textes doivent être normalisés pour minimiser l'influence de ces mots sur leur analyse. En effet, l'étape de normalisation transforme une copie d'un texte original dans un format standard plus facilement manipulable.

Pour l'étape de normalisation de textes, nous appliquons un certain nombre de traitements (e.g. le codage, la normalisation, etc.) pour nettoyer les textes des erreurs typographiques et réalisons un quelques opérations de normalisation afin de préparer le texte pour l'étape d'analyse. En effet, le codage vise à convertir les passages post-traités en codage UTF-8. Ce problème survient pendant le traitement de lettres arabes par le codage par défaut (e.g. le code ASCI). Par conséquent, avec le langage XML, un encodage UTF-8 devrait être spécifié dans le prologue du document XML : <?xml version="1.0" encoding="UTF-8"?>.

Considérant que, la normalisation vise à éliminer les données indésirables telles que mots vides, des chiffres et des signes de ponctuation [Sheker et al., 2016]. C'est une technique notable dans leur analyse. Elle se compose de plusieurs étapes de prétraitement, qui comprennent la suppression des translittérations, la suppression de la ponctuation (les points-virgules (;), colons (:), les points d'exclamation (!), les points d'interrogation (?), les traits d'union (-), les apostrophes (‘), les points de suspension (...), ..), l'élimination des diacritiques, le remplacement de أ et إ , آ avec ا, le filtrage des lettres non-arabes, etc.

2.3.3 Segmentation des passages

Les textes extraits du Web sont souvent longs, leur traitement consiste également à les découper en des phrases. De même, pour répondre à des questions, il est préférable d'éviter un long texte et extraire quelques phrases de ce texte qui peuvent être la réponse souhaitée. En effet, l'étape d'extraction de phrases de textes est une sorte de « Tokenization ». Elle sert à transformer le texte source écrit en arabe en une liste des phrases de taille plus au moins proche qui lui compose. Par conséquent, la segmentation de textes en phrases, paragraphes, items, etc., [Mourad, 2001], [Mouelhi 2008] reste une phase nécessaire et incontournable pour un très grand nombre d'applications en traitement automatique des langues. Néanmoins, la segmentation de textes arabes est toujours différente. Cela est du des particularités de cette langue dont il n'y a pas de majuscules qui indiquent le début d'une nouvelle phrase. Ainsi, les signes de ponctuation, ne sont pas utilisés de façon régulière. D'ailleurs, la segmentation de textes arabes peut être supportée aussi par des particules et certains mots tels que les conjonctions de coordination (e.g. "لكن " (lakin), "لقد " (laqad) et " أمّا " (amma)) ainsi que celles de certaines particules tels que les conjonctions de coordination ("و " (wa) et "ف " (f ā)) [Belguith et al., 2005].

2.3.4 Traitement linguistique des passages

Le traitement linguistique se compose d'un pipeline de traitement de langage naturel à usage général qui expose les différentes étapes pour pouvoir analyser un passage de texte généré à partir du Web. Cette analyse augmente la chance de trouver la réponse précise à une question en langage naturel.

- **La reconnaissance des entités nommées**

Après la segmentation des textes en phrases, l'étape suivante est augmentée par une étape représentant toutes les entités nommées dans le texte. Pour le faire, nous avons utilisé l'outil ArNER pour extraire des entités importantes telles que les noms, les organisations et les lieux qui sont ensuite enregistrées dans un fichier XML. La figure 4.17 montre le résultat de la reconnaissance des entités nommées d'un texte enrichi par les balises de segmentation en phrases d'un passage collecté à partir du Web pour la question suivante « **متى ولد حنبعل؟** ». Notons que certaines entités ne sont pas détectées selon l'outil ArNER.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <text>
  - <p>
    - <s num="1">
        <w num="1" type="B-PERS">حنّبعل</w>
        <w num="2" type="O">هو</w>
        <w num="3" type="O">من</w>
        <w num="4" type="O">أعظم</w>
        <w num="5" type="O">القادة</w>
        <w num="6" type="O">العسكريين</w>
        <w num="7" type="O">الذين</w>
        <w num="8" type="O">عرفهم</w>
        <w num="9" type="O">التاريخ</w>
        <w num="10" type="O">ولد</w>
        <w num="11" type="O">بقرطاج</w>
        <w num="12" type="O">ستة</w>
        <w num="13" type="O">247</w>
        <w num="14" type="O">قبل</w>
        <w num="15" type="O">الميلاد</w>
      </s>
    </p>
  </text>
```

Figure 4.17: Exemple d'un fichier XML émis par ArNER

- **L'analyse syntaxique**

Dans le cas de passages de textes, il s'agit de déterminer si les dépendances extraites et les phrases segmentées peuvent nous aider à construire une représentation sémantique correcte via les graphes conceptuels. Pour cela, nous accomplissons une analyse syntaxique en profondeur de la définition en utilisant le Stanford Parser. Cet analyseur donne une sortie sous forme de dépendances syntaxiques, comme montré dans la figure 4.18.

Figure 4.18: Sorties de Stanford pour le passage du texte P1

L'analyse morphologique

Dans le cas où les termes qui n'existent pas dans AWN, nous recherchons leurs tiges pour les ajouter à la liste de concepts. Le processus consiste à réduire les mots dérivés ou infléchis à leurs tiges ou à leurs racines originales. En utilisant Khoja Stemmer, chaque terme dans le texte d'entrée est représenté par sa tige et sa racine. Le terme «tige» a deux significations inconsidérément différentes. Tout d'abord, une tige peut être la partie centrale d'un mot qui exprime le sens de base et ne peut pas être divisé en plus petits morphèmes [Payne & Reader, 2006]. La figure 4.19 montre l'analyse morphologique d'un texte enrichi par les balises de segmentation en phrases. Ainsi, elle fournit une description détaillée de l'analyseur et sa sortie.

Figure 4.19: Sortie de Khoja stemmer pour le passage du texte P1

2.4 Représentation logique

Dans nos travaux, nous représentons le sens des questions et des passages réponses en des formes logiques pour réaliser une sorte d'implication entre elles. Afin d'établir quelle représentation logique est la mieux adaptée à chaque type spécifique de la question ou du

passage du texte, nous proposons de représenter la question et les passages en des graphes conceptuels. Par la suite, nous transformons ces graphes en des représentations logiques via le principe de l'opérateur Φ de Sowa. Enfin, nous déterminons la relation d'implication textuelle entre ces représentations, le chemin du travail est signalé par la figure 4.20.

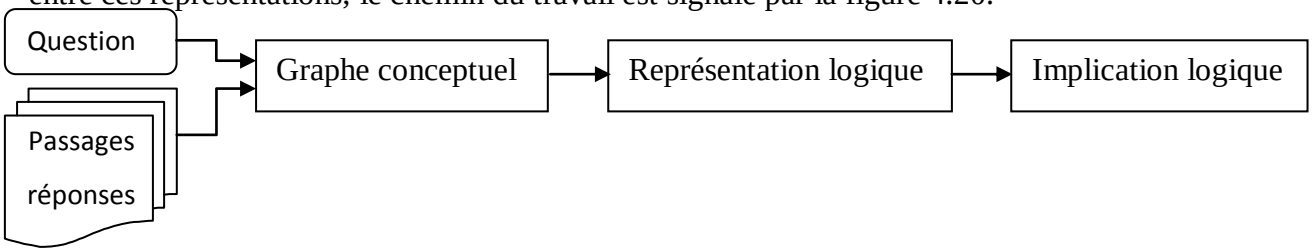


Figure 4.20: Etape de l'implication logique en un système de question-réponse Arabe

2.4.1 Construction d'une représentation sémantique avec les graphes conceptuels

Cette section présente notre méthode proposée pour la construction d'un graphe conceptuel à partir d'un texte arabe (question, passage). En effet, une modélisation conceptuelle en TALN est une façon de modéliser la sémantique. La sémantique des textes se transforme en sémantique de modèles conceptuels à un haut niveau d'abstraction, en termes de concepts et de relations. Par ailleurs, un concept représente un objet d'intérêt appelé aussi objet de connaissance. Les relations conceptuelles permettent d'associer les concepts. Un graphe conceptuel est un ensemble connexe de concepts et de relations conceptuelles. La transformation d'un texte (passage, question) en un graphe conceptuel est réalisée par 4 phases. D'abord, nous introduisons la liste des termes. Puis, nous détaillons l'étape d'extraction de concepts associés à ces termes. Ensuite, nous extrayons la liste de relations entre ces concepts. Enfin, nous construisons le graphe correspondant à ces concepts et relations.

a) Extraction de termes

Cette étape consiste à extraire, tout d'abord, les termes à partir des textes en arabe (question ou passages de textes) déjà pré-traités et analysés. En effet, ces termes sont des unités textuelles simples qui se comportent un texte, comme par exemple noms, verbes, adjectifs, etc. La technique adoptée pour l'identification de ces termes est d'éliminer la liste des mots vides. Ces derniers représentent les expressions les plus communes découvertes dans n'importe quelle langue naturelle et qui portent très peu ou pas de contexte sémantique expressif dans une phrase. Les exemples suivants sont des mots vides présentés en langue

arabe: **(‘ف’, ‘منذ’, ‘أو’,’هو’, ‘هناك’, ‘مهما’, etc.)**. Dans ce sens, la liste des mots vides est collectée du lien de cette source[17].

L’exemple du passage indiqué dans la figure 4.21 se compose de 2 phrases principales, dont le nombre total de mots, y compris les mots vides, est égal à 29 mots.

حنّبعل هو من أعظم القادة العسكريين الذين عرفهم التاريخ ولد بقرطاج سنة247 قبل الميلاد.“
ورافق وهو في التاسعة من عمره والده عبد ملقرط في حملته على إسبانيا“

Figure 4.21: Extrait d'un passage P1 de texte

Après avoir supprimé les mots vides et considéré ensuite les mots restants comme termes, le passage précédent contient 23 termes. Ces derniers sont enregistrés en un fichier XML, comme le montre la figure 4.22.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <Text>
  - <Sentence>
      <Term Value="حنبعل" pos="NN" />
      <Term Value="اعظم" pos="JJR" />
      <Term Value="القادة" pos="DTNN" />
      <Term Value="العسكريين" pos="DTJJ" />
      <Term Value="عرف" pos="VBD" />
      <Term Value="هم" pos="PRP" />
      <Term Value="التاريخ" pos="DTNN" />
      <Term Value="ولد" pos="VBN" />
      <Term Value="قرطاج" pos="DTNN" />
      <Term Value="سنة" pos="NN" />
      <Term Value="247" pos="CD" />
      <Term Value="الميلاد" pos="DTNN" />
    </Sentence>
  - <Sentence>
      <Term Value="" pos="NNP" />
    </Sentence>
  - <Sentence>
      <Term Value="رافق" pos="VBD" />
      <Term Value="التاسعة" pos="ADJ_NUM" />
      <Term Value="عمر" pos="NN" />
      <Term Value="ه" pos="PRP$" />
      <Term Value="والد" pos="NN" />
      <Term Value="ه" pos="PRP$" />
      <Term Value="عبد" pos="NNP" />
      <Term Value="ملقرط" pos="NNP" />
      <Term Value="حملة" pos="NN" />
      <Term Value="ه" pos="PRP" />
      <Term Value="اسبانيا" pos="NNP" />
    </Sentence>
</Text>
```

Figure 4.22: Liste de termes extraits d’un passage de texte

b) Extraction de concepts

Le principal challenge dans la construction des graphes conceptuels à partir des textes est l’identification automatique des concepts et des relations qui les relient [Rao et al., 2013]. Ainsi, nous avons élaboré une méthodologie qui utilise la reconnaissance d’entités nommées, l’analyse morphologique et l’analyse syntaxique pour extraire des concepts à partir des questions et des passages. Nous extrayons les concepts correspondants à tous les termes des

[17] http://www.ranks.nl/stopwords/arabic

textes en recourant sur WordNet arabe. Comme l'illustre la figure 4.23, les concepts d'un document en entrée sont extraits suivant les étapes suivantes:

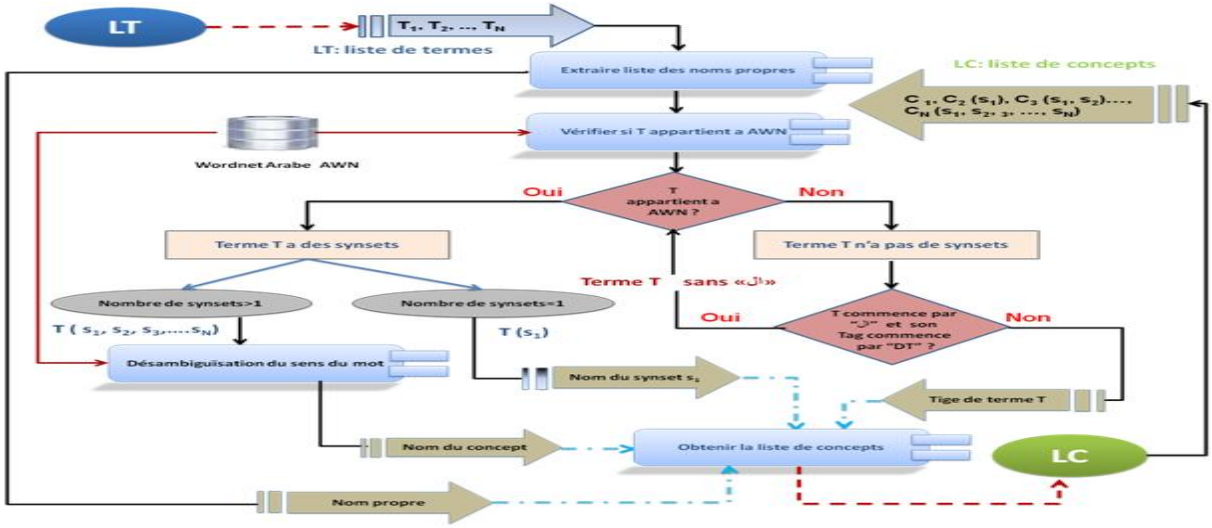


Figure 4.23: Extraction de concepts

- **Étape 1: Identifier les noms propres.**

Cette tâche est définie comme générique puisque tous les textes font usage de noms propres et que leur repérage semble a priori reproductible. Il est dès lors primordial de pouvoir reconnaître ces noms propres puisque WordNet arabe ne contient pas les noms propres. Généralement, trois types de noms propres ont été identifiés: les noms de personnes, les noms de lieux (localisations) et les noms d'organisations. Dans notre cas, l'identification se base essentiellement sur les noms de personnes. Par conséquent, les termes qui expriment des noms propres seront conservés tels quels sont et seront ajoutés à la liste des concepts extraits LC. Ces termes sont extraits en se basant sur le résultat de l'analyse syntaxique fournie par l'analyseur Stanford. Plus précisément, si le tag d'un terme T est égal à « NNP » alors c'est un nom propre.

- **Étape 2: Extraire les synsets de chaque terme**

Cette étape consiste à projeter la liste de termes retenus dans l'étape d'extraction de termes à WordNet arabe (AWN), afin de retourner les synsets qui leurs convient. En effet, AWN manipule les unités lexicales non pas par des mots mais par un ensemble de synonymes appelés « synset ». En outre, un synset est un ensemble de mots qui sont propriétaires de la même signification dans au moins un contexte. À partir de ce point, les termes sont classés, selon leur appartenance à AWN, en deux taxonomies possibles : des termes qui n'appartiennent pas à AWN, ils n'ont aucun synset et des termes qui appartiennent à AWN, ils font parti à différents synsets (1 ou plusieurs).

Dans le premier cas, l'API[18] JAVA pour AWN que nous l'employons afin d'accéder à sa base de données XML ne couvre pas tous les mots en arabe qui débute par « ال » par exemple, nous ne pouvons pas déterminer les synsets des mots comme «البرج, القسم » ce qui est possible avec «برج , قسم ». Pour cette raison, dans le cas des mots qui débutent par « ال » et ont le Tag commence par « DT » nous proposons d'éliminer « ال » et de projeter de nouveau ces termes à AWN. En revanche, si les termes ne débutent pas par « ال » et n'ayant pas un Tag commencé par « DT », nous allons extrayons leurs tiges et nous les ajoutons à la liste des concepts extraits LC.

Dans le second cas, pour associer des synsets aux concepts, deux stratégies peuvent être déterminées: des termes avec un seul synset et des termes possèdent plusieurs synsets. En ce qui concerne les termes qui font parti à un seul synset, nous allons extraire le nom du concept et nous l'ajoutons à la liste des concepts extraits LC. Enfin, les termes qui appartiennent à plus qu'un synset sont appelés des termes ambigus. Dans ce cas, nous proposons d'attribuer à chaque mot ambigu un sens approprié. Cela sera bien décrit dans l'étape de « **Désambiguïsation du sens du mot ».**

Tableau 4.16: exemple de synsets des termes : « البرج » et « برج» à partir d'AWN

Terme	Synsets
البرج	aucun
برج	[بَرْج, سَفِيْنَة],[بُرجْ]

- **Étape 3: Désambiguïsation du sens du mot**

Chaque mot, à partir du contenu du texte, peut appartenir à un ou plusieurs sens. Cela entraînera une ambiguïté dans l'analyse de son contenu. L'identification computationnelle de la signification pour les mots dans le contexte est appelée désambiguïsation des sens des mots (WSD). Cette tâche permet d'identifier le sens correct d'un mot ambigu dans un contexte donné [Navigli, 2009]. En effet, la langue arabe est connue comme une langue riche sémantiquement, un mot peut avoir plusieurs sens selon leur contexte d'utilisation. À cet effet, la désambigüisation devient une tâche importante afin de lever l'ambigüité des mots en question. Pour cela, nous présentons une méthode de désambiguïsation du sens du mot qui combine le WordNet arabe [Black et al., 2006], l'étiqueteur automatique Stanford POS Tagger et un dictionnaire arabe-arabe «**المعجم الوسيط**» [Muṣṭafá et al., 2008] qui contient les différentes définitions de ce mot ambigu, en tant que ressources de désambiguïsation. Enfin, nous appliquons l'algorithme du Lesk simplifié pour distinguer le sens exact des différents

[18] https://sourceforge.net/projects/javasourcecodeapiarabicwordnet/

sens donnés [Lesk, 1986]. Cet algorithme représente une méthode de désambiguïsation bien connue qui consiste à compter le nombre de mots communs entre les définitions des mots de son contexte et les définitions d'un mot (généralement trouvées dans un dictionnaire électronique).

Dans ce qui suit, nous décrivons brièvement le déroulement de l'étape qui permet de trouver les sens corrects des termes ambigus qui sont déjà identifiés par une méthode de désambiguïsation sémantique que nous proposons. Ce déroulement sera schématisé dans la figure 4.24. Notre processus de WSD a été accompli en effectuant les étapes suivantes:

- **Etape 3.1:** Déterminer tous les mots ambigus possibles en utilisant WordNet arabe.
- **Etape 3.2 :** Réduire le nombre des sens, en se basant sur l'étiqueteur morpho-syntaxique Stanford POS Tagger, et laisser seulement les sens ayant le même Tag que le terme. Pour cela chaque phrase d'entrée est transmise au Stanford POS Tagger pour obtenir les étiquettes morpho-syntaxiques de chaque mot dans la phrase.
- **Etape 3.3 :** Présenter toutes les définitions et les exemples correspondants à leurs synsets d'un dictionnaire pour chaque mot ambigu selon leur Tag (par exemple seulement les définitions et les exemples d'un nom si le mot ambigu est un nom, ou d'un verbe si le mot ambigu est un verbe).
- **Etape 3.4 :** Appliquer l'algorithme Lesk simplifié afin d'extraire le sens approprié.

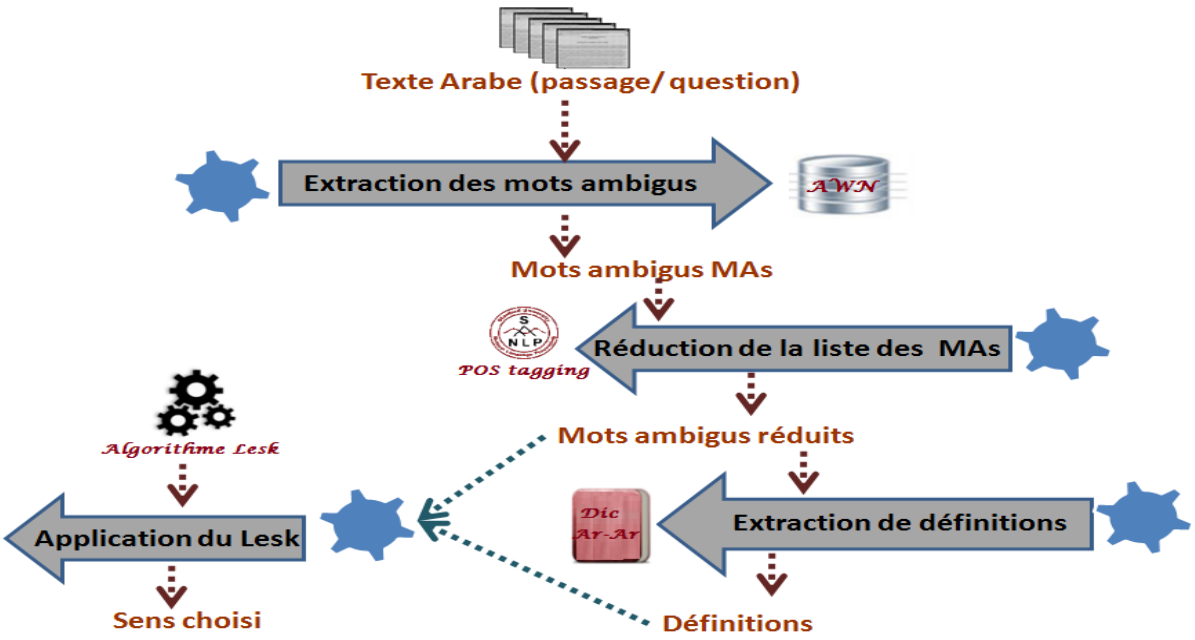


Figure 4.24: Schéma descriptif de désambiguïsation avec l'algorithme de Lesk

✓ **Exemple de désambiguïsation avec l'algorithme de Lesk**

Dans ce qui suit, nous présentons un exemple décrivant le principe de la méthode de désambiguïsation utilisée : En entrée, nous utilisons la phrase **Ph_2** « **ورافق وهو في التاسعة من عمره والده عبد ملقرط في حملته على إسبانيا**» du passage de la figure 4.21. En sortie l'algorithme produit les sens les plus appropriés aux termes à désambiguïser, présentés dans le tableau 4.17.

Tableau 4.17: Sens choisis par Lesk de la phrase Ph_2

Terme ambigu	Synsets sélectionnés à partir d'AWN ayant le même Tag que le terme ambigu	Sens choisi
رافق	{رافق, لازم, صحب} {رافق, لازم, كمن} {رافق} {رافق, لازم, صاحب} {رافق, صاحب}	**رافق**
عمر	{مدة حياة, حياة, عمر, فترة حياة} {حياة, عمر} {سن, عمر} {سن, عمر}	**عمر**
والد	{منجب, أب, والد}	**منجب**
حملة	{حركة, حملة} {حملة, حملة عسكرية} {حملة انتخابية, حملة} {حملة إنتخابية, حملة سياسية, حملة}	**حملة**

- **Étape 4: Obtenir la liste des concepts**

Un graphe conceptuel est composé de deux types de nœuds, des nœuds concepts représentant les entités et des nœuds de relation représentant les relations entre ces entités. Les concepts peuvent être définis comme des significations ou des idées ayant différents niveaux sémantiques derrière des termes spécifiques dans un document texte [Bleik et al., 2010] ; ils sont dénotés dans des documents XML. Les concepts et les relations sont utilisés pour créer une représentation avec le formalisme d'un graphe conceptuel.

Figure 4.25: Concepts associés aux termes de la figure 4.22

c) Extraction de relations

Une fois que tous les concepts ont été extraits à partir d'un texte, toutes les relations entre les paires de concepts doivent être ainsi récupérées. Pour le faire, nous avons besoin des Tags aussi que des dépendances de chaque concept. En effet, les dépendances fournies de Stanford procurent une représentation des relations grammaticales entre les mots dans une phrase. Elles ont étés conçues pour être facilement comprises et utilisées efficacement par des personnes qui souhaitent extraire des relations textuelles. En outre, les dépendances de Stanford sont des triplets: la dépendance, le gouverneur et le dépendant. Par exemple, soit la dépendance: **nsubj (ولد-11, حنبعل-1)** :

- Le type de dépendance est « **nsubj** »
- Le dépendant est **« حنبعل»**
- Le gouverneur est **« ولد »**

Dans ce cadre, nous extrayons toutes les dépendances trouvées par Stanford. D'abord, pour chaque dépendance nous identifions le type de dépendance, le dépendant et le gouverneur. Puis, nous cherchons le tag du dépendent « Dtag » et le tag du gouverneur « GTag ». Ensuite, à partir de la liste des concepts LCs, nous vérifions si le mot a un concept ; si oui, selon le type de dépendance, le tag du dépendent, le tag du gouverneur, nous choisissons la règle adéquate parmi les règles proposées par [Abouenour, 2014]. Finalement, cette règle est appliquée pour extraire la relation associée à ces types de concepts. Les règles proposées par Abouenour sont mentionnées dans l'annexe B. Pour les questions et les passages réponses, toutes les 11 règles ont été appliquées au moins une fois. Un exemple d'application de ces règles est rapporté dans le tableau 4.18. En appliquant les règles de Abouenour, nous trouvons dans ce cas la règle adéquate pour trouver la relation adéquate est la règle 1.

Tableau 4.18: principe de la règle 1 de [Abouenour, 2014]

Règle	**Exemple**
Règle 1: "GTag=JJ and DTag=NN" If the Governor(head) Tag (GTag) is "JJ" and the Dependent Tag (DTag) is a noun, then there are two cases: The dependent tag is neither "NNP" nor "NNPS": in this case the conceptual graph of the dendency is constructed following the pattern: ***CG-dep = [cg : [Conc(G)]<-attributeOf-[Conc(D)]]***	**[عظم]**<-(attributeOf)-**[قود]**

Selon l'analyse syntaxique fournie par Stanford, nous obtenons:

Tableau 4.19: Analyse de dépendance fournie par Stanford

dependant	Head	type_dep	tag_dependant	tag_head
القادة	اعظم	dep	**DTNN**	**JJR**

Ensuite, nous transformons le pattern en remplaçant Conc(G) par « **عظم**» et Conc(D) par « **قود**», et nous construisons enfin le graphe conceptuel sous la forme linéaire suivante :

[عظم]<-(attributeOf)-**[قود]**

Finalement, une fois tous les concepts et toutes les relations entre eux ont été extraits d'un texte, ils sont représentés avec le formalisme des graphes conceptuels.

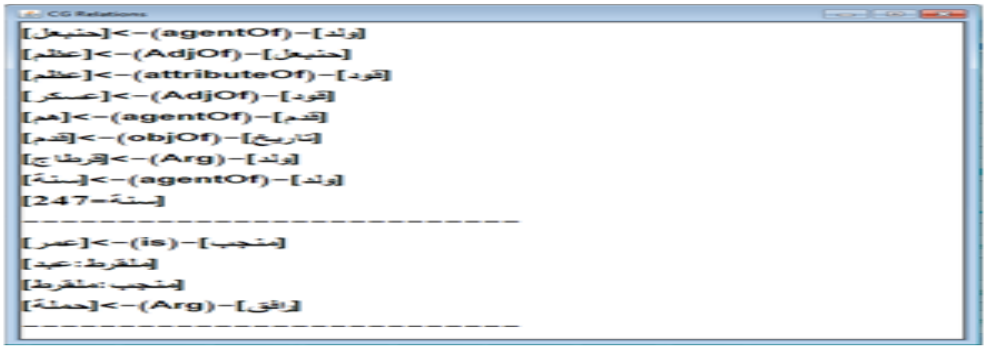


Figure 4.26: Liste de relations retenues du passage P1 en appliquant les règles d'Abouenour

d) Construction du graphe

La constitution d'un graphe conceptuel des textes en arabe (texte ou question) suit la démarche illustrée dans la figure 4.27. Chacune de ces étapes a été bien décrite en détails dans les sections supérieures. En effet, la représentation d'un document en arabe (question et passages) via des graphes conceptuels est effectuée pour déterminer la représentation logique de ce document.

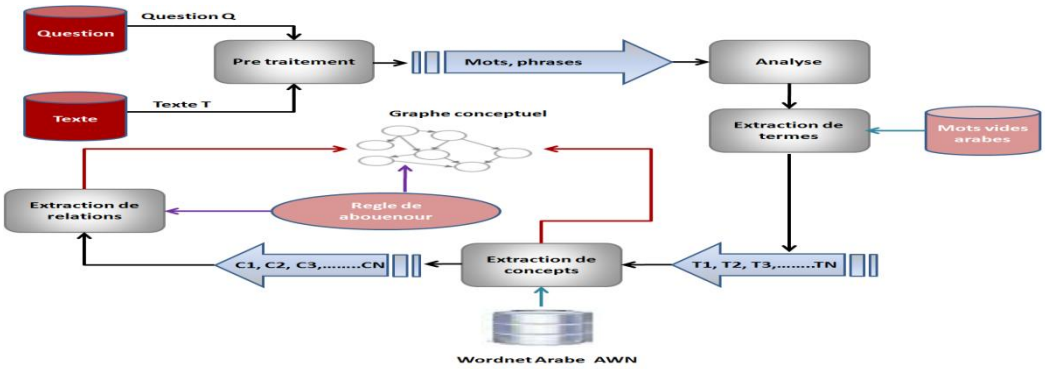


Figure 4.27: Construction d'un graphe conceptuel

e) Représentation des pharses d'un passage avec le formalisme des graphes conceptuels

Actuellement, chaque phrase d'un texte est représentée par un graphe conceptuel.

❖ **Cas d'une phrase verbale**

Tableau 4.20: Représentation d'une phrase verbale avec le formalisme des graphes conceptuels

Exemple 1 : (Verbe intransitif) « نزل المطر »		
Listes de termes	**Listes de concepts**	**Listes de relations et graphe correspondant**
<?xml version="1.0" encoding="UTF-8" ?> - <Text> - <Sentence> <Term Value="نزل" pos="VBD" /> <Term Value="المطر" pos="DTNN" /> </Sentence> </Text>	<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <AllConcepts> - <Sentence> <Concept>نزل</Concept> <Concept>مطر</Concept> </Sentence> </AllConcepts>	[نزل]<-(objOf)-[مطر] ----------
Exemple 2 : (Verbe transitif) « أكمل التلميذ الدرس »		
Listes de termes	**Listes de concepts**	**Listes de relations et graphe correspondant :**
<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <Text> - <Sentence> <Term Value="اكمل" pos="VBD" /> <Term Value="التلميذ" pos="DTNN" /> <Term Value="الدرس" pos="DTNN" /> </Sentence> </Text>	<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <AllConcepts> - <Sentence> <Concept>كمل</Concept> <Concept>تلميذ</Concept> <Concept>درس</Concept> </Sentence> </AllConcepts>	[تلميذ]<-(agentOf)-[كمل] [كمل]<-(objOf)-[درس]

❖ **Cas d'une phrase nominale**

Tableau 4.21: Représentation d'une phrase nominale avec le formalisme des graphes conceptuels

Exemple 3 : phrase nominale avec un attribut NG : « السماء زرقاء »		
Listes de termes	**Listes de concepts**	**Listes de relations et graphe correspondant :**
<?xml version="1.0" encoding="UTF-8" ?> - <Text> - <Sentence> <Term Value="السماء" pos="DTNN" /> <Term Value="زرقاء" pos="JJ" /> </Sentence> </Text>	<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <AllConcepts> - <Sentence> <Concept>زرق</Concept> <Concept>سماء</Concept> </Sentence> </AllConcepts>	*Relations :* [زرق]<-(attributeOf)-[سماء] ---------- *Graphe :* [سماء]- -(attributeOf)->[زرق] ----------
Exemple 4 : phrase nominale avec un attribut VG : أحمد ذهب إلى الجامعة (2)		
Listes de termes	**Listes de concepts**	**Listes de relations et graphe correspondant :**

<?xml version="1.0" encoding="UTF-8" ?> - <Text> - <Sentence> <Term Value="احمد" pos="NNP" /> <Term Value="ذهب" pos="VBD" /> <Term Value="الجامعة" pos="DTNN" /> </Sentence> </Text>	<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <AllConcepts> - <Sentence> <Concept>احمد</Concept> <Concept>ذهب</Concept> <Concept>جامعة</Concept> </Sentence> </AllConcepts>	***Relations :*** [Person: احمد]<-(agentOf)-[ذهب] [جامعة]<-(Arg)-[ذهب] --- ***Graphe :*** [ذهب]- -(agentOf)->[Person: احمد] -(Arg)->[جامعة] ---

f) Représentation des questions avec le formalisme de graphes conceptuels

En réalité, la représentation d'une question via le formalisme d'un graphe conceptuel se diffère de celle d'une phrase d'un passage et nécessite une attention particulière. Par conséquent, la représentation avec le formalisme d'un graphe conceptuel a été effectuée comme s'il s'agissait d'une phrase déclarative, puis les informations inconnues sont remplacées par des variables (e.g. la question «أين تقع تونس؟» est considérée comme «Location : LX تقع تونس في »).

Tableau 4.22: Liste de termes, concepts et relations des exemples de questions

Exemple 1 : (Question «من »)		
Question : من اخترع الحاسوب ؟ (1)		
Déclaration de la Question : الحاسوب **Person: PX** اخترع		
Listes de termes	**Listes de concepts**	**Listes de relations**
<?xml version="1.0" encoding="UTF-8" standalone=" - <Question> - <Terms> <Term Value="من" pos="IN" /> <Term Value="اخترع" pos="DTNN" /> <Term Value="الحاسوب" pos="DTJJ" /> </Terms> </Question>	<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <Question> - <Concepts> <Concept>من</Concept> <Concept>اخترع</Concept> <Concept>حاسوب</Concept> </Concepts> </Question>	[Person:PX]<-(agentOf)-[اخترع] [اخترع]<-(objOf)-[حاسوب]
Exemple 2 : (Question «أين»)		
Question : أين تقع تونس ؟ (2)		
Déclaration de la Question : **Location: LX** في تونس تقع		
Listes de termes	**Listes de concepts**	**Listes de relations**
<?xml version="1.0" encoding="UTF-8" ?> - <Question> - <Terms> <Term Value="اين" pos="WRB" /> <Term Value="تقع" pos="VBP" /> <Term Value="تونس" pos="NNP" /> </Terms> </Question>	<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <Question> - <Concepts> <Concept>اين</Concept> <Concept>تقع</Concept> <Concept>تونس</Concept> </Concepts> </Question>	[Location:LX]<-(Loc)-[تقع] [NE:تونس]<-(objOf)-[تقع]
Exemple 3 : (Question «ماهو»)		
Question : ماهو أكبر تنظيم اقتصادي في العالم؟ (3)		
Déclaration de la Question : **Organisation : OX** هو أكبر تنظيم اقتصادي في العالم		
Listes de termes	**Listes de concepts**	**Listes de relations**

<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <Question> - <Terms> <Term Value="ماهو" pos="VBD" /> <Term Value="أكبر" pos="JJR" /> <Term Value="تنظيم" pos="NN" /> <Term Value="اقتصادي" pos="JJ" /> <Term Value="العالم" pos="DTNN" /> </Terms> </Question>	<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <Question> - <Concepts> <Concept>ماهو</Concept> <Concept>كبر</Concept> <Concept>تنظيم</Concept> <Concept>اقتصادي</Concept> <Concept>العالم</Concept> </Concepts> </Question>	[Organisation:OX]<-(is)-[كبر] [كبر]<-(attributeOf)-[تنظيم] [اقتصادي]<-(propertyOf)-[تنظيم] [Organisation:OX]<-(is)-[العالم]

Exemple 4 : (Question «متى»)

Question : متى تأسست حركة حماس؟ **(4)**

Déclaration de la Question : **Date : TX** تأسست حركة حماس في

Listes de termes	Listes de concepts	Listes de relations
<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <Question> - <Terms> <Term Value="متى" pos="WRB" /> <Term Value="تاسست" pos="VBD" /> <Term Value="حركة" pos="NN" /> <Term Value="حماس" pos="NNP" /> </Terms> </Question>	<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <Question> - <Concepts> <Concept>متى</Concept> <Concept>أسس</Concept> <Concept>حركة</Concept> <Concept>حماس</Concept> </Concepts> </Question>	[date:TX]<-(TMP)-[متى] [أسس]<-(objOf)-[حركة] [حركة:حماس]

Exemple 5 : (Question «كم»)

Question : كم عدد الممالك في أوروبا؟ **(5)**

Déclaration de la Question : **Numerical expression : NX** عدد الممالك في أوروبا

Listes de termes	Listes de concepts	Listes de relations
<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <Question> - <Terms> <Term Value="كم" pos="WRB" /> <Term Value="عدد" pos="NN" /> <Term Value="الممالك" pos="DTNN" /> <Term Value="أوروبا" pos="NNP" /> </Terms> </Question>	<?xml version="1.0" encoding="UTF-8" standalone="no" ?> - <Question> - <Concepts> <Concept>كم</Concept> <Concept>عدد</Concept> <Concept>ملك</Concept> <Concept>أوروبا</Concept> </Concepts> </Question>	[Numerical_expression:NX]<-(Value)-[كم] [أوروبا]<-(attributeOf)-[ملك] [عدد:أوروبا]

Tableau 4.23: Graphes conceptuels correspondants aux exemples des questions du tableau 4.22

Question	Type de la question	Graphe conceptuel
(1) من اخترع الحاسوب ؟	Personne	[اخترع]– –(agentOf)–>[Person:PX] [حاسوب]– –(objOf)–>[اخترع]
(2) أين تقع تونس ؟	Location	[تقع]– –(Loc)–>[Location:LX] –(objOf)–>[NE:تونس]
(3) ماهو أكبر تنظيم اقتصادي في العالم؟	Organisation	[كبر]– –(is)–>[Organisation:OX] [تنظيم]– –(attributeOf)–>[كبر] –(propertyOf)–>[اقتصادي] [العالم]– –(is)–>[Organisation:OX]
(4) متى تأسست حركة حماس؟	Date	[متى]– –(TMP)–>[date:TX] [حركة]– –(objOf)–>[أسس] [حركة:حماس]
(5) كم عدد الممالك في أوروربا؟	Expression numérique	[كم]– –(Value)–>[Numerical_expression:NX] [ملك]– –(attributeOf)–>[اوروبا] [عدد:اوروبا]

2.4.2 Raisonnement logique à l'aide des graphes conceptuels

Un graphe conceptuel peut également être construit comme une structure logique d'un certain type. En effet, Sowa [Sowa, 1984] a proposé plusieurs règles fondamentales permettent de manipuler de manière cohérente les graphes conceptuels. En particulier, une caractéristique importante de leur modèle est que les raisonnements présentés sur les graphes peuvent être faits tout en conservant un lien avec la logique de premier ordre. Dans nos travaux, nous décrivons notre algorithme de transformation des graphes conceptuels qui est reposé sur le principe de l'opérateur Φ proposé dans [Sowa, 1984] afin d'obtenir une représentation en logique du premier ordre des informations récupérées de ces graphes de la question et des passages réponses.

Définition : (*L'opérateur Φ*) : L'association d'une formule logique Φ(u) à un graphe conceptuel u.

Algorithme **Conversion_Graphe_Conceptuel_à_Formule_Logique_Premier_Ordre**

1 : ***VARIABLES***
2 : GC :Graphe /* Graphe Conceptuel(Entrée)
3 : C :LISTE /* Liste pour les concepts.
4 : R :LISTE /* Liste pour les relations.
5 : Conjonction :CHAINE
6 : Symbole :CHAINE
7 : FLPO :LISTE /* Formule Logique de Premier Ordre(Sortie)
8 : Prédicati :CHAINE
9 : Pcj :CHAINE
10 : Prédicatj :CHAINE
11 : Prédicatr :CHAINE
12 : Pci :CHAINE
13 :**DEBUT_ALGORITHME**
14 : FLPO=Null,C=Null,R=Null,Symbole= «∃ »,Conjonction= « ∧»
15 : ExtractionConcept(GC,C),ExtractionRelation(GC,R)
16 : P =C,Q= P^suivant
17 : **Tantque**(P<>Null) **Faire**
18 : **SI** (Marqueur_Individuel (P^Concept)=vrai et Marqueur_Individuel (Q^Concept)=vrai)) **ALORS**
19 : Pci =constantei
20 : Prédicati =P^TypeConcept(Pci)

21 : Pcj =une variable indépendante et quantifiée
22 : Prédicatj =Q^TypeConcept(Pcj)
23 : Prédicatr =R^TypeRelation(Pci, Pcj,)
24 : InsertTete(symbole, Pcj, FLPO)
25 : InsertQueue(Prédicati, Conjonction, Prédicatr , Conjonction, Prédicatj, FLPO)
26 : **FIN_SI**
27 : **SI** (Marqueur_Individuel (P^Concept)=faux et Marqueur_Individuel (Q^Concept)=faux) **ALORS**
28 : Pci =une variable indépendante et quantifiée
29 : Pcj =une variable indépendante et quantifiée
30 : Prédicati = P^TypeConcept(Pci)
31 : Prédicatj =Q^TypeConcept(Pcj)
32 : Prédicatr =R^TypeRelation(Pci, Pcj)
33 : InsertTete(symbole, Pci, FLPO)
34 : InsertTete(symbole, Pcj, FLPO)
35 : InsertQueue(Prédicati, Conjonction, Prédicatr , Conjonction, Prédicatj, FLPO)
36 : **FIN_SI**
37 : **SI** (Marqueur_Individuel (P^Concept)=faux et Marqueur_Individuel (Q^Concept)=vrai) **ALORS**
38 : Pci = une variable indépendante et quantifiée
39 : Pcj = constantej
40 : Prédicati = P^TypeConcept(Pci)
41 : Prédicatj =Q^TypeConcept(Pcj)
42 : Prédicatr =R^TypeRelation(Pci, Pcj)
43 : InsertTete(symbole, Pci, FLPO)
44 : InsertQueue(Prédicati, Conjonction, Prédicatr , Conjonction, Prédicatj, FLPO)
45 : **FIN_SI**
46 : P=P^suivant, Q= P^suivant, R=R^suivant
47 : **FINTantque**
48 : **FIN_ALGORITHME**

Un concept est caractérisé par un type et éventuellement un marqueur : le type représente la classe sémantique à laquelle un objet appartient, le marqueur permet de nommer et de distinguer les différents objets d'une classe. Par contre, une relation est caractérisée par un type. L'algorithme proposé transforme chaque élément dans un graphe conceptuel en un élément de la logique du premier ordre. Il associe à tout type de concept un prédicat et à tout type de relation, un prédicat de même arité que le type. Il associe également aux marqueurs individuels des constantes et aux marqueurs génériques des variables qui peuvent être

considérées comme arguments. Par conséquent, une interprétation en forme logique des graphiques conceptuels avec un quantificateur universel et existentiel a été donnée. Dans nos travaux, les expressions logiques générées sont simplifiées. En effet, les pluriels, les modalités et les quantifications ne sont pas prises en compte ; les conditionnels et les négations sont représentés comme des prédicats normaux et non comme opérateurs logiques. Finalement, le résultat est une formule logique de premier ordre obtenue par conjonction des prédicats associés aux concepts et aux relations.

2.4.3 Transformation des graphes conceptuels en des représentations logiques

En utilisant l'algorithme proposé pour la transformation d'un graphe conceptuel en une forme logique, nous obtenons pour chaque graphe donné sa représentation en logique de premier ordre correspondante. En général, la transformation d'une phrase en langue naturelle (question, passage) dans la logique de premier ordre sert d'abord à écrire le prédicat, puis à associer leurs arguments correspondants.

a) Cas des questions

La représentation logique tente de saisir le sens de la question [Nyberg et al., 2003]. Cette procédure est répétée pour différents exemples du même type de la question.

Tableau 4.24: Représentations logiques des graphes conceptuels des questions du tableau 4.23

Les mots de la question	Représentation logique de la question
Question	*(1) من اخترع الحاسوب ؟*
من *اخترع* *الحاسوب*	∃X ∃Y : Person(PX) ∧ اخترع(X) ∧ agentOf(X,PX) ∧ حاسوب(Y) ∧ objOf(Y,X)
Question	*(2) أين تقع تونس ؟*
أين *تقع* *تونس*	∃X ∃Y : Location(LX) ∧ قعي(X) ∧ Loc(X,LX) ∧ NE(تونس) ∧ قعي(Y) ∧ objOf(Y,تونس)
Question	*(3) ماهو أكبر تنظيم اقتصادي في العالم؟*
ماهو *أكبر* *تنظيم* *اقتصادي* *في* *العالم*	∃X ∃Y ∃Z ∃W : Organisation(OX) ∧ كبر(X) ∧ is(X,OX) ∧ تنظيم(Y) ∧ attributeOf(Y,X) ∧ اقتصادي(Z) ∧ propertyOf(Y,Z) ∧ العالم(W) ∧ is(W,OX)
Question	*(4) متى تأسست حركة حماس؟*
متى *تأسست* *حركة* *حماس*	∃X ∃Y ∃Z ∃W : date(TX) ∧ متى(X) ∧ TMP(X,TX) ∧ حركة(Y) ∧ أسس(Z) ∧ objOf(Y,Z) ∧ حماس(W) ∧ is(Y,W)
Question	*(5) كم عدد الممالك في أوروربا؟*

كم عدد الممالك في أوروبا	∃X ∃Y ∃Z ∃W : Numerical_expression(NX) ∧ كم(X) ∧ Value(X,NX) ∧ ملك(Y) ∧ اوروبا(Z) ∧ attributeOf(Y,Z) ∧ عدد(W) ∧ is(W,Z)

b) Cas des passages

Dans cette section, la structure des formes logiques et leur génération à partir des passages est décrite pour tout type de phrase. Pour les phrases verbales nous avons défini la transformation pour la classe des verbes transitifs et celle des verbes intransitifs. De ce fait, la question est de savoir quels arguments sont impliqués dans n'importe quelle situation est déterminée par la signification du prédicat.

(1) نزل المطر

Prenant l'exemple de la phrase (1) qui traite un cas d'une phrase verbale dont le verbe est intransitif. La conversion de cette phrase en sa représentation logique correspondante sera comme suit :

$$\exists X\ \exists Y : \text{مطر}(X) \wedge \text{نزل}(Y) \wedge objOf(X,Y)$$

(2) أكمل التلميذ الدرس

Prenant l'exemple de la phrase (2) qui traite un cas d'une phrase verbale dont le verbe est transitif. La conversion de cette phrase en sa représentation logique correspondante sera comme suit :

$$\exists X\ \exists Y\ \exists Z : \text{كمل}(X) \wedge \text{تلميذ}(Y) \wedge agentOf(X,Y) \wedge \text{درس}(Z) \wedge objOf(Z,X)$$

Les phrases nominales sont généralement composées d'un topique (Mubtada') et d'un attribut (Khabar). Ces derniers se présentent sous plusieurs formes qui peuvent modifier la structure de la phrase. Le topique peut être un nom propre, un pronom personnel, un nom et un adjectif, une préposition et un nom, un nom suivi par une conjonction de coordination suivie d'un autre nom ou un nom suivi d'un déterminant suivi d'un pronom attaché. L'attribut peut être un nom, un adjectif, un pronom personnel, un nom et un adjectif, un nom suivi d'un autre nom défini, une préposition et un nom ou bien tout un syntagme verbal. Dans nos travaux, nous ne prenons pas en compte les différents types du topique mais nous étudions les types de l'attribut en les regroupant en deux classes.

La première classe appelée groupe nominal (NG) rassemble les types : un nom, un adjectif, un pronom personnel, un nom et un adjectif, un nom suivi d'un autre nom défini, une

préposition et un nom. Prenant l'exemple de la phrase (3), la conversion de cette phrase en sa représentation logique correspondante sera comme suit :

(3) السماء زرقاء

$$\exists X \exists Y : \text{سماء}(X) \wedge \text{زرق}(Y) \wedge attributeOf(X,Y)$$

La deuxième classe appelée groupe verbal (VG) présente l'attribut sous forme d'un syntagme verbal. Prenant l'exemple de la phrase (4), la conversion de cette phrase en sa représentation logique correspondante sera comme suit :

(4) أحمد ذهب إلى الجامعة

$$\exists X \exists Y : Person(\text{احمد}) \wedge \text{ذهب}(X) \wedge agentOf(X,\text{احمد}) \wedge \text{جامعة}(Y) \wedge Arg(X,Y)$$

2.4.4 Détermination de l'implication textuelle

Étant donné que notre approche déploie une analyse sémantique et logique peu profonde. D'un point de vue logique, la preuve d'une implication textuelle consiste à montrer qu'une représentation logique est déductible d'une ou plusieurs d'autres représentations. Nous proposons une méthode qui est essentiellement basée sur l'extraction et la combinaison des caractéristiques pour déterminer la relation d'implication logique entre chaque paire d'une question et son passage réponse.

a) Etape 1 : Extraction des caractéristiques

Dans cette étape, nous découvrons trois caractéristiques utilisées pour déterminer l'implication textuelle. Étant donné que seules les caractéristiques peu profondes sont prises en compte dans nos travaux. Ces caractéristiques ont été construites pour être utilisées dans la détermination, la classification des implications et la sélection de la réponse précise. Les caractéristiques sont basées sur des scores calculés par une métrique de similarité. Pour chaque métrique, nous utilisons les représentations logiques d'entrée des paires T-H (passage réponse- question) respectivement FOLT et FOLH ; les caractéristiques sont les suivantes:

(i) Caractéristique 1 : Chevauchement des Prédicats-arguments

Dans nos travaux, pour mesurer le chevauchement des mots entre le texte et l'hypothèse, nous supposons que nous parlons des entités similaires. Notre modèle général est un simple modèle de sac à mots. Donc, le chevauchement est généré pour toutes les paires de mots (w 1, w 2) où w 1 ∈ FOLT et w 2 ∈ FOLH. Les mots peuvent être soit ***un prédicat unaire,*** soit ***un prédicat binaire***. En effet, un prédicat binaire consiste en une paire

d'arguments qui sont liés pouvant représenter une constante ou une variable qui se réfère à un prédicat. Cependant, un prédicat unaire consiste d'un seul argument qui est soit une constante soit une variable. Dans le cas d'un prédicat binaire, nous prenons en compte les relations (e.g: ***"objOf(X,تونس)***"). Finalement, dans le cas d'un prédicat unaire, si l'argument représente alors une constante (e.g. ***"تونس"***), nous prenons son nom pour la comparaison. Par contre, si l'argument est une variable (e.g : ***"X"***), nous prenons en compte son prédicat correspondant. Le chevauchement est calculé comme suit :

$$\text{Predicat}_{\text{argument overlap}} = \frac{\text{Nombre de mots en commun entre le texte et l'hypothèse}}{\text{Nombre total de mots dans l'hypothèse}}$$

(ii) Caractéristique 2 : Correspondance entre entités nommées

Pour chaque entité nommée NE1 dans la représentation logique du passage nous cherchons alors une entité nommée NE2 dans l'hypothèse qu'elle implique. Par conséquent, nous considérons que NE1 de FOLT implique NE2 de FOLH si la chaîne de texte de NE1 contient la chaîne de texte de NE2. Plus précisément, le processus de correspondance entre les entités nommées suit les démarches suivantes :

1. Extraire les listes d'entités nommées trouvées dans FOLT, FOLH et dans le fichier résultat d'ArNER.
2. Comparer chaque entité nommée de FOLT avec les entités nommées de FOLH :
   i. Soit :
      - **NEP** : Nombre de NE partagées entre FOLH et FOLT, est initialisé à 0.
      - **NNEH** : Nombre d'entités nommées de FOLH.
   ii. Pour chaque ENH de FOLH de type (Personne, Organisation, Location, etc.), nous cherchons les ENT de FOLT avec le même type.
   iii. S'il se produit dans FOLT une ENT de même type qui correspond à ENH, alors la variable **NEP** est incrémentée.
      - Si la chaîne de texte d'ENT contient la chaîne de texte d'ENH alors ENT correspond à ENH.
      - Si non, nous calculons la mesure de similarité d(NET,NEH) en utilisant la distance de Levenshtein. Si les deux chaînes sont différentes à moins de 20% alors NET correspond à ENH. Cette distance est calculée comme suit :

$$\textbf{Distance Levenshtein} = \frac{\text{distLev}}{\text{maxLength}} * 100$$

Avec : distLev = La distance de Levenshtein.

maxLength = la longueur de la chaine de caractère entre NE1 et NE2 ayant longueur maximale.

3. Calculer le score de correspondance $\mathbf{Score_{NE_{corresp}}}$ entre FOLH et FOLT. Si, pour une paire FOLT/FOLH, le score est inférieur à 1, cela indique qu'il existe au moins une entité nommée dans FOLH qui ne se produit pas dans FOLT. Par conséquent, nous pouvons déduire que l'implication est fausse. Sinon, l'implication est vraie. Le score de correspondance est calculé comme suit :

$$\mathbf{Score_{NE_{corresp}}} = \frac{\text{Nombre de NE partagées entre le texte et l'hypothèse}}{\text{Nombre des NEs dans l'hypothèse}}$$

(iii) Caractéristique 3 : Similarité sémantique

La caractéristique de similarité sémantique permet de fournir un score de similarité entre deux représentations logiques de la question et de chaque passage réponse. Ce score est calculé à partir des relations sémantiques entre les différents mots en utilisant le WordNet arabe. En se basant sur ce score, nous pouvons décider si les deux formes logiques sont liées ou non, et ainsi déduire l'implication entre le texte (passage) et l'hypothèse (question).

Parfois, nous trouvons des mots qui sont différents mais qui expriment la même signification ou inversement. Notons, que ces mots ne sont pas traités avec la première caractéristique où le chevauchement des mots peut être bien établi si tous les mots dans l'hypothèse sont exactement appariés dans le texte. Pour cela, nous suggérons de rechercher la similarité sémantique entre chaque mot de l'hypothèse avec tous les mots du texte en utilisant le WordNet arabe. D'ailleurs, dans WordNet arabe, chaque mot est organisé en des taxonomies où chaque nœud est un ensemble de synonymes (synset) représentés par un sens. Si un mot a plus d'un sens, il apparaîtra dans plusieurs synsets à différents endroits de la taxonomie. WordNet arabe définit les relations entre les synsets et les relations entre les sens des mots. Une relation entre synsets est une relation sémantique, et une relation entre les sens des mots est une relation lexicale. La différence est que les relations lexicales sont des relations entre les membres de deux synsets différents, mais les relations sémantiques sont des relations entre deux synsets entiers. (Par exemple hypernym, hyponym, holonym, etc., et les relations lexicales sont la relation d'antonyme et la relation de la forme dérivée).

Pour déterminer la similarité sémantique entre deux représentations logiques, nous nous cherchons, en premier lieu, la similarité sémantique entre chaque mot de l'hypothèse avec tous les mots du texte en utilisant la mesure de parenté sémantique Wu-Palmer fournie par AWN. En effet, Wu-Palmer est fondée sur les profondeurs et les longueurs dans la taxonomie. Elle prend en considération la longueur entre les concepts C1 et C2, la longueur entre le LCS et la racine de la classification dans laquelle les concepts existent. La similarité est déterminée comme suit:

$$sim_{WP}(C_i, C_j) = \frac{2 * \text{depth}(\text{lCS}(C_i, C_j))}{depth(C_i) + depth(C_j)}$$

Avec :

- Depth(C) est la profondeur du synset C en utilisant le comptage des arêtes dans la taxonomie.
- LCS (C1, C2) est le plus petit sous-segment en commun de C1 et C2.
- Depth (LCS (C1, C2)) est la longueur entre LCS de C1 et C2 et la racine de la taxonomie.

Puis, nous calculons la similarité globale entre deux représentations logiques en utilisant la stratégie moyenne correspondante. Etant donné les deux formes logiques FOLH et FOLT qui correspondent respectivement à l'hypothèse et à un passage réponse quelconque, nous désignons m pour la longueur de FOLH, n pour longueur de FOLT. Pour calculer la moyenne correspondance, nous proposons de construire une matrice relative de similarité sémantique R [m, n] de chaque paire de sens des mots de FOLH et FOLT, où R [i, j] est la similarité sémantique entre le mot à la position i de FOLH et le mot à la position j de FOLT. Ainsi, la similarité entre FOLH et FOLT est réduite au problème d'un calcul de poids d'appariement total maximum d'un graphe bipartite. Ceci est assuré en exploitant l'algorithme hongrois sur ce graphe où X et Y sont FOLH et FOLT et les nœuds du graphe sont les mots apparentés [Dao & Simpson, 2005]. La correspondance moyenne calculée comme suit:

$$Moyenne\_Correspondante = \frac{2 * Match(FOLH, FOLT)}{|FOLH| + |FOLT|}$$

b) Etape 2 : Détermination de l'implication

Étant donné pour chaque question (hypothèse) et leurs passages réponses (texte) qui sont représentés en des formes logiques, trois caractéristiques intéressantes peuvent être répertoriées, à savoir, le chevauchement des prédicats-arguments, la correspondance des

entités nommées et la similarité sémantique, indiquant la similitude entre eux. Dans nos travaux, la sélection des caractéristiques est essentielle pour la détermination de l'implication.

- *Exemple1 : déterminer l'implication avec le chevauchement des Prédicats-arguments*

Commençons par un premier exemple pour comprendre le principe de la caractéristique de chevauchement des prédicats-arguments.

Question : أين يوجد مقر منظمة اليونسكو ؟

FOLH : ***∃LX ∃X ∃Y ∃Z ∃W : Location(LX) ∧ يوجد(X) ∧ Loc(X,LX) ∧ مقر(Y) ∧ objOf(Y,X) ∧ منظمة(Z) ∧ is(Z,Y) ∧ اليونسكو(W) ∧ is(Z,W)***

Passage réponse : يوجد مقر منظمة اليونسكو في باريس بفرنسا

FOLT : ***∃X ∃Y ∃Z ∃W ∃T ∃E : مقر(X) ∧ يوجد(Y) ∧ objOf(X,Y) ∧ منظمة(Z) ∧ is(Z,X) ∧ اليونسكو(W) ∧ is(Z,W) ∧ باريس(T) ∧ is(Z,T) ∧ فرنسا(E) ∧ Arg(Y,E)***

Dans cet exemple, nous calculons d'abord le chevauchement entre les paires des prédicats unaires. Donc, cet exemple comporte quatre paires se chevauchent (e.g. ceux écrits en bleu). Nous évaluons leur chevauchement comme suit:

$$\text{Unary}_{\text{predicate overlap}} = \frac{\text{NombreTnH}}{\text{NombreH}} = \frac{4}{5} = 0.80$$

Maintenant, nous calculons les chevauchements entre les paires des prédicats binaires. Nous présentons, d'abord, l'ensemble des relations de FOLH par R1= ***{"Loc(X,LX)", "objOf(Y,X)", "is(Z,Y)" et "is(Z,W)"}***, l'ensemble des relations de FOLT par R 2= ***{" objOf(X,Y)", " is(Z,X)", " is(Z,W)", "is(Z,T)" et "Arg(Y,E)" }***. Nous constatons que la paire « FOLH/FOLT » comporte trois relations partagées, leur chevauchement est égal :

$$\text{Binary}_{\text{predicate overlap}} = \frac{\text{NombreTnH}}{\text{NombreH}} = \frac{3}{4} = 0.75$$

Maintenant, nous calculons la mesure de chevauchement « Predicat_argument_overlap» des prédicats unaires et binaires, elle est égale à :

$$\text{Predicat}_{\text{argument overlap}} = \frac{\text{Unary}_{\text{predicate overlap}} + \text{Binary}_{\text{predicate overlap}}}{2} = 0.77$$

- *Exemple 2 : Déterminer l'implication avec la correspondance des ENs*

Une fois que les entités nommées de la paire de FOLT-FOLH ont été détectées, l'étape suivante consiste à déterminer les relations d'implication entre elles en se basant sur la

caractéristique de correspondance. Dans certains cas, les entités nommées ne peuvent pas être distingues parce que leurs types ne sont pas décrits dans la forme logique. Pour résoudre ce souci, nous prenons en considération le résultat d'ArNER et vérifions s'il existe d'autres entités nommées pour déterminer le score de correspondance.

Question : أين يوجد مقر منظمة اليونسكو ؟

FOLH : $\exists LX\ \exists X\ \exists Y\ \exists Z\ \exists W : Location(LX) \wedge$ وجد$(X) \wedge Loc(X,LX) \wedge$ مقر$(Y) \wedge objOf(Y,X) \wedge$ منظمة$(Z) \wedge is(Z,Y) \wedge$ اليونسكو$(W) \wedge is(Z,W)$

Passage réponse : يوجد مقر منظمة اليونسكو في باريس بفرنسا

FOLT : $\exists X\ \exists Y\ \exists Z\ \exists W\ \exists T\ \exists E :$ مقر$(X) \wedge$ وجد$(Y) \wedge objOf(X,Y) \wedge$ منظمة$(Z) \wedge is(Z,X) \wedge$ اليونسكو$(W) \wedge is(Z,W) \wedge$ باريس$(T) \wedge is(Z,T) \wedge$ فرنسا$(E) \wedge Arg(Y,E)$

Dans cet exemple, nous calculons le score d'appariement entre FOLT et FOLH. Donc, cet exemple comporte une paire des entités nommées qui se partagent, à savoir, «Organisation « منظمة اليونسكو », le score attribué est déterminé comme suit :

$$\mathbf{Score}_{\mathbf{NE_{corresp}}} = \frac{1}{2} = 0.5$$

Notons que, dans certains cas, l'implication en utilisant la correspondance entres NE1 et NE2 est déterminée mais pas complète. Où parfois, dans les paires où il y a correspondance entre entités, il n'y a pas assez d'informations pour décider si la valeur d'implication est vraie. Cependant, dans d'autres cas, la correspondance ne peut pas être assurée (à cause des résultats produits par ArNER). Egalement, certains caractères changent dans différentes expressions de la même entité nommée que, par exemple, dans un nom propre possédant des mots variés (p.e.x. « فرانكلين», «فرانكلن », « فرنكلن »). Une même entité nommée peut contenir des formes d'orthographe différentes, d'abréviations et aussi de variantes similaires, mais non identiques.

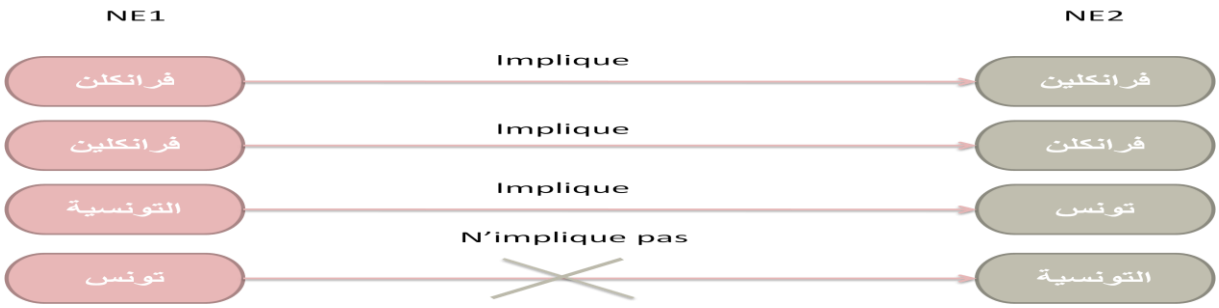


Figure 4.28: Des exemples d'implication entre entités nommées

Pour résoudre ces problèmes, une mesure de similarité efficace devrait être déterminée si deux entités partagent la même entité ou non. Cette similarité est adoptée en utilisant la distance de Levenshtein comme étant une mesure simple de la distinction entre deux chaînes de texte. Cette distance calcule le nombre de changements (basé sur le caractère) pour générer une chaîne de texte à partir de l'autre. Dans cette distance, nous considérons les mots comme les plus petites unités lors du calcul. Par conséquent, dans le cas d'une paire de FOLT/FOLH, si l'hypothèse contient une entité nommée qui ne se trouve pas dans le texte, le texte ne peut pas impliquer l'hypothèse.

Nous avons trouvé, en se recourant à la distance de Levenshtein, que les deux chaînes des entités nommées NE1= **فرانكلن** et NE2= **فرانكلين** se diffèrent de 12.5% qui est inférieure à 20%, alors nous considérons que la correspondance entre NE1 et NE2 est vraie. Pour NE1= **التونسية** et NE2= **تونس**, nous avons une NE1 contient NE2, alors NE1 implique NE2 mais NE2 n'implique pas NE1.

Exemple 3 : Déterminer l'implication avec la similarité sémantique

Considérons un exemple d'une hypothèse et d'un passage de texte avec leurs représentations logiques :

Question : أين يوجد مقر منظمة اليونسكو ؟

FOLH : $\exists LX\ \exists X\ \exists Y\ \exists Z\ \exists W : Location(LX) \wedge$ وجد$(X) \wedge Loc(X,LX) \wedge$ مقر$(Y) \wedge objOf(Y,X) \wedge$ منظمة$(Z) \wedge is(Z,Y) \wedge$ اليونسكو$(W) \wedge is(Z,W)$

Passage réponse : يوجد مقر منظمة اليونسكو في باريس بفرنسا

FOLT : $\exists X\ \exists Y\ \exists Z\ \exists W\ \exists T\ \exists E :$ مقر$(X) \wedge$ وجد$(Y) \wedge objOf(X,Y) \wedge$ منظمة$(Z) \wedge is(Z,X) \wedge$ اليونسكو$(W) \wedge is(Z,W) \wedge$ باريس$(T) \wedge is(Z,T) \wedge$ فرنسا$(E) \wedge Arg(Y,E)$

Nous cherchons d'abord la similarité sémantique entre les synsets de tous les mots de FOLH avec chaque mot de FOLT en utilisant la méthode Wu-Palmer. Nous construisons ensuite la matrice relative de la similarité sémantique de chaque paire de sens des mots. Les résultats de correspondance sont combinés alors en une seule valeur de similarité par la correspondance moyenne pour la paire FOLT et FOLH qui est égale à **0.72**.

c) Etape 3 : Combinaison des caractéristiques

Nous avons identifié trois caractéristiques pertinentes pour la détermination de l'implication textuelle. Nous analysons ainsi comment une combinaison de caractéristiques

pourrait avoir un impact sur la performance globale de la décision d'implication. Plus précisément, trois caractéristiques correspondantes sont combinées pour juger la relation d'implication au lieu d'une seule. Nous avons choisi seulement les trois caractéristiques mentionnées précédemment car en tenant compte du fait que de plus grands ensembles de caractéristiques ne conduisent pas obligatoirement à l'amélioration des performances de la décision d'implication. Ainsi, un grand nombre de caractéristiques pourrait augmenter le risque de donner de fausses informations ou de perdre des passages qui sont susceptibles de comporter la réponse. D'ailleurs, nous trouvons dans certains cas qu'il est difficile de trouver un passage qui détermine en même temps toutes les caractéristiques malgré qu'il contienne une réponse à la question donnée. Par conséquent, nous arrivons à déterminer que la combinaison de caractéristiques est souvent plus performante que la meilleure caractéristique individuelle dans l'ensemble.

d) Etape 4 : Classification des implications

La dernière étape consiste à effectuer une classification des implications. A un certain niveau, le problème d'implication est simplement considéré comme un problème de classification. Plus précisément, c'est un problème à deux classes, une classe « OUI » et une classe « NON ». Dans nos travaux, pour chaque hypothèse (question) et leurs passages réponses respectivement représentés en des formes logiques, FOLH et FOLTs, nous utilisons les caractéristiques qui ont été évaluées séparément d'abord comme des données d'apprentissage pour former des vecteurs caractéristiques. Ces derniers ont transmis à un classificateur d'arbre de décision J48 de WEKA [Witten & Frank, 1999] qui les classe en tant que des implications « OUI » ou « NON ». Le raison principale de l'utilisation du classificateur d'arbre de décision est qu'il est simple à interpréter, ce qui nous a permis d'apprendre quelles sont les caractéristiques les plus discriminatoires et lesquelles sont les moins.

Nous avons classé 250 paires texte-hypothèse dans la tâche d'implication textuelle ; le texte est un ensemble de passages qui peuvent contenir une réponse à une question (hypothèse). Un vecteur de caractéristiques est construit pour ces exemples en utilisant les scores qui sont fournis par chaque caractéristique. Ce vecteur est utilisé à la fois pour la classification et pour l'association des scores de confiances aux implications. Dans la classification, ce vecteur est utilisé dans le processus d'apprentissage automatique qui comporte principalement deux étapes telles que l'entraînement et le test. Parmi les exemples

traités, nous utilisons 2/3 pour la phase d'entraînement et 1/3 pour la phase de test. Enfin, le classificateur génère à la fois une relation d'implication textuelle (oui ou non) aussi un score de confiance. Le tableau 4.25 montre un exemple de paires de FOLH-FOLTs avec leurs relations d'implication fournis par le classificateur d'arbre de décision. Nous discutons avec ces exemples qu'il existe 4 passages qui ont une relation d'implication avec l'hypothèse et 9 passages n'ayant pas.

Tableau 4.25: Résultat de la classification des implications entre FOLH-FOLTs

	FOLH : ∃LX ∃X ∃Y ∃Z ∃W : Location(LX) ∧ وجد(X) ∧ Loc(X,LX) ∧ مقر(Y) ∧ objOf(Y,X) ∧ منظمة(Z) ∧ is(Z,Y) ∧ اليونسكو(W) ∧ is(Z,W)	**C1**	**C2**	**C3**	**Implication**
FOLT1	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H ∃I : أسس(X) ∧ منظمة(Y) ∧ agentOf(X,Y) ∧ لمم(Z) ∧ attributeOf(Z,Y) ∧ حدد(W) ∧ AdjOf(Z,W) ∧ تربية(T) ∧ attributeOf(T,Y) ∧ معرفة(E) ∧ attributeOf(E,T) ∧ ثقافة(F) ∧ attributeOf(F,T) ∧ اليونسكو(G) ∧ is(Y,G) ∧ حول(H) ∧ objOf(H,X) ∧ 1945(I) ∧ isEqual(H,I)	**0.32**	**0.5**	**0.28**	**NON**
FOLT2	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F : مقر(X) ∧ قعي(Y) ∧ objOf(X,Y) ∧ منظمة(Z) ∧ is(Z,X) ∧ اليونسكو(W) ∧ is(Z,W) ∧ باريس(T) ∧ Arg(Y,T) ∧ جذر(E) ∧ Arg(Y,E) ∧ حديث(F) ∧ propertyOf(E,F)	**0.55**	**0.5**	**0.5**	**OUI**
FOLT3	∃X ∃Y ∃Z ∃W ∃T : كبي(X) ∧ اليونسكو(Y) ∧ Arg(X,Y) ∧ ميد(Z) ∧ objOf(Z,X) ∧ انحاء(W) ∧ Arg(X,W) ∧ العالم(T) ∧ attributeOf(T,W)	**0.10**	**0.0**	**0.25**	**NON**
FOLT4	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H : مقر(X) ∧ وجد(Y) ∧ objOf(X,Y) ∧ منظمة(Z) ∧ is(Z,X) ∧ اليونسكو(W) ∧ is(Z,W) ∧ باريس(T) ∧ Arg(Y,T) ∧ مقر(E) ∧ objOf(E,Y) ∧ روس(F) ∧ AdjOf(E,F) ∧ منظمة(G) ∧ is(G,E) ∧ اليونسكو(H) ∧ is(G,H)	**0.77**	**0.5**	**0.57**	**OUI**
FOLT5	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H ∃I ∃J : وكالة(X) ∧ اليونسكو(Y) ∧ is(X,Y) ∧ خصص(Z) ∧ propertyOf(X,Z) ∧ منظمة(W) ∧ تتبع(T) ∧ objOf(W,T) ∧ لمم(E) ∧ attributeOf(E,W) ∧ حدد(F) ∧ AdjOf(E,F) ∧ قوم(G) ∧ agentOf(T,G) ∧ تبع(H) ∧ Arg(G,H) ∧ دولة(I) ∧ 191(J) ∧ isEqual(I,J) ∧ objOf(I,G)	**0.20**	**0.0**	**0.25**	**NON**
FOLT6	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H ∃I ∃J ∃K ∃L ∃M ∃N : ملك(X) ∧ منظمة(Y) ∧ agentOf(X,Y) ∧ اليونسكو(Z) ∧ is(Y,Z) ∧ رمج(W) ∧ خمس(T) ∧ isEqual(W,T) ∧ objOf(W,X) ∧ سوس(E) ∧ propertyOf(W,E) ∧ تربية(F) ∧ attributeOf(F,W) ∧ تعليم(G) ∧ attributeOf(G,F) ∧ علم(H) ∧ attributeOf(H,F) ∧ طبع(I) ∧ AdjOf(H,I) ∧ علم(J) ∧ attributeOf(J,F) ∧ أنس(K) ∧ propertyOf(J,K) ∧ ثقافة(L) ∧ attributeOf(L,F) ∧ اتصالات(M) ∧ attributeOf(M,F) ∧ علم(N) ∧ attributeOf(N,M)	**0.32**	**0.5**	**0.24**	**NON**
FOLT7	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H ∃I ∃J ∃K : غرض(X) ∧ روس(Y) ∧ AdjOf(X,Y) ∧ انشاء(Z) ∧ is(Z,X) ∧ منظمة(W) ∧ is(W,Z) ∧ اليونسكو(T) ∧ is(W,T) ∧ سهم(E) ∧ attributeOf(E,X) ∧ هدف(F) ∧ is(F,E) ∧ سلام(G) ∧ attributeOf(G,F) ∧ أمن(H) ∧ attributeOf(H,G) ∧ سلام(I) ∧ is(I,E) ∧ دول(J) ∧ is(J,I) ∧ العالم(K) ∧ attributeOf(K,J)	**0.32**	**0.5**	**0.28**	**NON**
FOLT8	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H ∃I ∃J ∃K ∃L ∃M ∃N ∃O : عرف(X) ∧ منظمة(Y) ∧ agentOf(X,Y) ∧ منظمة(Z) ∧ اليونسكو(W) ∧ is(Z,W) ∧ is(Z,Y) ∧ علم(T) ∧ propertyOf(Z,T) ∧ كلمة(E) ∧ agentOf(X,E) ∧ اليونسكو(F) ∧ is(E,F) ∧ خصر(G) ∧ agentOf(X,G) ∧ منظمة(H) ∧ objOf(H,X) ∧ منظمة(I) ∧ أصل(J) ∧ is(I,J) ∧ لمم(K) ∧ attributeOf(K,I) ∧ حدد(L) ∧ AdjOf(K,L) ∧ تربية(M) ∧ attributeOf(M,I) ∧ علم(N) ∧ attributeOf(N,M) ∧ ثقافة(O) ∧ attributeOf(O,M)	**0.32**	**0.5**	**0.23**	**NON**
FOLT9	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H ∃I ∃J : NE(اليونسكو) ∧ دعم(X) ∧ agentOf(X,اليونسكو) ∧ عديد(Y) ∧ objOf(Y,X) ∧ شرع(Z) ∧ attributeOf(Z,Y) ∧ محو(W) ∧ Arg(X,W) ∧ امية(T) ∧ attributeOf(T,W) ∧ تدريب(E) ∧ attributeOf(E,T) ∧ قنأ(F) ∧ AdjOf(E,F) ∧ رمج(G) ∧ is(G,W) ∧ أهل(H) ∧ is(H,G) ∧ تدريب(I) ∧ is(I,H) ∧ علم(J) ∧ attributeOf(J,H)	**0.10**	**0.0**	**0.22**	**NON**
FOLT10	∃X ∃Y ∃Z ∃W ∃T ∃E : مقر(X) ∧ وجد(Y) ∧ objOf(X,Y) ∧ منظمة(Z) ∧ is(Z,X) ∧ اليونسكو(W) ∧ is(Z,W) ∧ باريس(T) ∧ is(Z,T) ∧ فرنسا(E) ∧ Arg(Y,E)	**0.77**	**0.5**	**0.72**	**OUI**
FOLT11	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H ∃I : منظمة(X) ∧ نمي(Y) ∧ objOf(X,Y) ∧ اليونسكو(Z) ∧ is(X,Z) ∧ فصيلة(W) ∧ Arg(Y,W) ∧ لمم(T) ∧ attributeOf(T,W) ∧ حدد(E) ∧ AdjOf(T,E) ∧ شطر(F) ∧ ها(G) ∧ agentOf(F,G) ∧ مثل(H) ∧ objOf(H,F) ∧ هدف(I) ∧ is(I,H)	**0.32**	**0.5**	**0.31**	**OUI**
FOLT12	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H ∃I ∃J ∃K ∃L : وجد(X) ∧ منظمة(Y) ∧ agentOf(X,Y) ∧ لمم(Z) ∧ attributeOf(Z,Y) ∧ حدد(W) ∧ AdjOf(Z,W) ∧ تربية(T) ∧ attributeOf(T,Y) ∧ علم(E) ∧ attributeOf(E,T) ∧ ثقافة(F) ∧ attributeOf(F,T) ∧ خصر(G) ∧ عرف(H) ∧ objOf(G,H) ∧ اليونسكو(I) ∧ Arg(H,I) ∧ حلة(J) ∧ objOf(J,X) ∧ روس(K) ∧ propertyOf(J,K) ∧ باريس(L) ∧ Arg(X,L)	**0.30**	**0.0**	**0.38**	**NON**

FOLT13	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H ∃I ∃J ∃K ∃L ∃M ∃N : دشن(X) ∧ سال(Y) ∧ objOf(X,Y) ∧ جذر(Z) ∧ attributeOf(Z,X) ∧ روس(W) ∧ AdjOf(Z,W) ∧ واقع(T) ∧ attributeOf(T,Z) ∧ ساحة(E) ∧ Arg(Y,E) ∧ فوننتوا(F) ∧ is(E,F) ∧ باريس(G) ∧ Arg(Y,G) ∧ منزل(H) ∧ حضن(I) ∧ objOf(H,I) ∧ اليونسكو(J) ∧ is(H,J) ∧ نوفمبر(K) ∧ 3(L) ∧ isEqual(K,L) ∧ is(K,H) ∧ شري(M) ∧ is(M,K) ∧ 1958(N) ∧ isEqual(M,N)	**0.10**	**0.0**	**0.19**	**NON**

2.5 **Recherche de la réponse précise**

La recherche de la réponse précise a pour but de fouiller tous les passages pertinents pour sélectionner le passage qui contient une réponse exacte et précise à la question posée. Dans notre cas, l'étape d'extraction et de sélection d'une réponse peut être modélisée par une implication textuelle logique. D'ailleurs, les passages qui ont la valeur "vraie" dans l'étape de détermination d'implication sont sélectionnés comme des réponses candidates. Enfin, le passage ayant le score de confiance le plus élevé est décidé d'être « la bonne réponse ».

2.5.1 **Extraction et pondération des réponses candidates**

Après avoir reconnu l'implication textuelle entre les représentations logiques des passages et de la question, l'objectif de cette étape est double:

- Extraire seulement les passages ayant une implication textuelle à la question de l'utilisateur.
- Attribuer des scores de confiances à chacun de ces passages.

En d'autres termes, nous écartons tous les passages qui n'ont aucune relation d'implication avec la question. Un score de confiance à chaque implication est ensuite fourni. Pour obtenir ce score, nous suggérons de déterminer la moyenne pondérée. Dans ce cadre, les valeurs des caractéristiques indiquent l'ensemble de données et les poids représentent les évaluations de gain d'information de chaque attribut. Ce gain est calculé par Weka pour les exemples d'entraînement préparés de chaque caractéristique. Ainsi, le score de confiance est calculé par la formule suivante :

$$score\_de\_confiance = \frac{\sum_{i=1}^{n} \alpha_i \, f_i}{\sum_{i=1}^{n} \alpha_i}$$

Avec :

- f= $\{f_1, .. f_n\}$: l'ensemble des valeurs des caractéristiques.
- α= $\{\alpha_1, .., \alpha_n\}$: l'ensemble des gains de chaque caractéristique.
- n est le nombre de caractéristique qui est égal à 3 dans nos travaux.

Ce score de confiance est représenté par des nombres réels entre 0 et 1. Les passages sont ordonnés dans un ordre décroissant. Ensuite, le passage ayant le score de confiance le

plus élevé est a priori contient la réponse à la question. Ainsi, un nombre prédéfini de réponses de passages (ici des phrases) est finalement sélectionné en fonction de ces scores. Ces réponses sont stockées pour sélectionner la première phrase comme étant la réponse à la question. A titre d'exemple, le tableau 4.26 montre les passages retenus avec implication ordonnés en ordre décroissant selon leurs scores de confiance. En fonction de la valeur de ce score, nous décidons quel passage parmi ceux retenus avec relation d'implication est le plus approprié pour répondre à la question.

Tableau 4.26: Scores attribués aux passages retenus avec implication

	FOLH : ∃LX ∃X ∃Y ∃Z ∃W : Location(LX) ∧ وجد(X) ∧ Loc(X,LX) ∧ مقر(Y) ∧ objOf(Y,X) ∧ منظمة(Z) ∧ is(Z,Y) ∧ اليونسكو(W) ∧ is(Z,W)	**Implication**	**Score**	**Ordre**
FOLT10	**∃X ∃Y ∃Z ∃W ∃T ∃E : مقر(X) ∧ وجد(Y) ∧ objOf(X,Y) ∧ منظمة(Z) ∧ is(Z,X) ∧ اليونسكو(W) ∧ is(Z,W) ∧ باريس(T) ∧ is(Z,T) ∧ فرنسا(E) ∧ Arg(Y,E)**	**OUI**	**0.75**	**1**
FOLT4	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H : مقر(X) ∧ وجد(Y) ∧ objOf(X,Y) ∧ منظمة(Z) ∧ is(Z,X) ∧ اليونسكو(W) ∧ is(Z,W) ∧ باريس(T) ∧ Arg(Y,T) ∧ مقر(E) ∧ objOf(E,Y) ∧ روس(F) ∧ AdjOf(E,F) ∧ منظمة(G) ∧ is(G,E) ∧ اليونسكو(H) ∧ is(G,H)	OUI	0.65	2
FOLT2	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F : مقر(X) ∧ قعي(Y) ∧ objOf(X,Y) ∧ منظمة(Z) ∧ is(Z,X) ∧ اليونسكو(W) ∧ is(Z,W) ∧ باريس(T) ∧ Arg(Y,T) ∧ جذر(E) ∧ Arg(Y,E) ∧ حديث(F) ∧ propertyOf(E,F)	OUI	0.52	3
FOLT11	∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H ∃I : منظمة(X) ∧ نمي(Y) ∧ objOf(X,Y) ∧ اليونسكو(Z) ∧ is(X,Z) ∧ فصيلة(W) ∧ Arg(Y,W) ∧ لمم(T) ∧ attributeOf(T,W) ∧ حدد(E) ∧ AdjOf(T,E) ∧ شطر(F) ∧ ها(G) ∧ agentOf(F,G) ∧ مثل(H) ∧ objOf(H,F) ∧ هدف(I) ∧ is(I,H)	OUI	0.30	4

2.5.2 Sélection de la réponse précise

Dans cette étape, nous nous intéressons à l'extraction de la réponse exacte. Dans nos travaux, une réponse correcte est extraite comme un passage (une phrase). Plus précisément, lorsque touts les passages sont pondérés, nous devrions sélectionner le passage réponse qui semble le plus pertinent et correct en utilisant le score de confiance. De plus, parmi les n passages reconnus avec implication, la question nécessite, comme réponse, le passage de cette collection qui a le score de confiance le plus élevé. Pour la question **Q-exemple**, « أين يوجد مقر منظمة اليونسكو ؟ », nous déterminons d'abord les passages candidats comme réponses à cette question ; 4 passages sont retenus avec implication (e.g. FOLT2, FOLT4, FOLT10 et FOLT11). Enfin, la sélection d'une réponse précise consiste à choisir « FOLT10 » « يوجد مقر منظمة اليونسكو في باريس بفرنسا» comme le passage le plus pertinent qui contient une réponse précise à cette question.

Conclusion

Pour finir, ce chapitre contient une description détaillée de notre approche proposée pour répondre aux questions arabes. Cette approche est fondée sur l'idée de d'intégrer la notion de logique et d'inférence dans la question-réponse arabe afin d'obtenir une meilleure performance. Notamment, les approches à base de la logique et l'inférence sont des approches hybrides entre celles morphosyntaxiques et sémantiques. À l'heure actuelle, les approches fondées sur la logique et l'inférence sont peu étudiées en arabe. Dans nos travaux de recherche, nous avons montré que ce type d'approches est un sujet original pour la question-réponse arabe. Nous avons ainsi utilisé quelques outils de TALN pour l'analyse morphologique et la reconnaissance des entités nommées pour analyser la question et le passage de texte récupéré à partir du Web. Nous avons vu comment tester l'implication entre deux énoncés en langue naturelle en les transformant en des représentations logique, permettant ainsi de générer des couples de textes s'impliquant ou non.

Dans le chapitre suivant, nous présentons l'évaluation de notre approche, aussi bien en termes de performances d'analyse de la question et des passages réponse qu'en termes d'extraction d'une réponse correcte pour assurer la satisfaction des utilisateurs. Notre approche est implémentée en un système; nous présentons également les principales spécifications de ce système ainsi que son architecture et son fonctionnement.

CHAPITRE 5 : DEVELOPPEMENT ET EVALUATION D'UN SYSTEME DE QUESTION-REPONSE POUR LA LANGUE ARABE

Conclusion…….. 144

1. Description générale 144

1.1 Présentation de NArQAS 145

1.2 Outils, ressources et techniques utilisés 146

2. Conception et implémentation de NArQAS 150

2.1 Architecture de NArQAS 150

2.2 Vue de Fonctionnement 152

2.3 Composants de NArQAS : entrées-sorties 153

2.4 Détails de l'implémentation 157

3. Evaluation et résultats obtenus 162

3.1 Ensemble de données pour l'évaluation 163

3.2 Mesures d'évaluation utilisées 163

3.3 Résultats obtenus par NArQAS 164

4. Analyse des résultats expérimentaux 166

4.1 Cas d'erreurs et traitements d'amélioration 166

4.2 Performance de NArQAS en comparaison avec Qwant et Ask.com 171

Conclusion…….. 175

Introduction

Nous présentons dans ce chapitre la structure générale de notre système de génération de réponses précises à des questions en arabe nommé NArQAS (New Arabic Question Answering System). Nous présentons l'évaluation des sorties de chacun de ces composants en se basant sur une collection de questions et de textes récupérés à partir du Web. Ce système a pour but de développer et d'évaluer l'apport de l'utilisation des procédures de raisonnement sémantique-logique, des techniques de traitement de langues naturelles ainsi que la technologie RTE afin d'élaborer des réponses précises à des questions factuelles. Nous détaillons également les principaux objectifs de NArQAS ainsi que son architecture de fonctionnement. Notamment, notre système est vu comme comme un apport, plutôt qu'un rival, aux systèmes classiques focalisés sur des approches extensivement basées sur des techniques de recherche d'informations et des techniques de TALN. Nous concluons ce chapitre en comparant notre système à d'autres travaux similaires.

1. Description générale

L'intérêt de notre approche a été illustré à travers le système «NArQAS». Ce système a fait l'objet d'une évaluation montrant que la prise en compte de nouveaux types d'approches notamment, à base de la sémantique et la logique permet d'améliorer la question-réponse arabe ainsi d'améliorer les résultats des différents modules du processus de traitement. En effet, NArQAS est un système complet allant de l'analyse de la question en langue naturelle à la génération de réponses en langue naturelle. C'est un système hybride combinant un analyseur sémantique avec des raisonnements logique; il comporte principalement cinq étapes, telles que l'analyse de la question, la récupération de passages, la représentation logique des énoncés, la détection des implications textuelles entre une question et un passage de texte et l'extraction de la réponse. En générale, chaque étape possède un besoin particulier d'informations en entrée et exécute des actions sur les informations extraites pour produire des résultats.

Notre système est basé sur un raisonnement logique sémantique efficace pour l'analyse des questions et des passages et des métriques d'implication textuelle pour l'extraction et la sélection de la réponse exacte. Ainsi, NArQAS, est un système modulaire, dont leurs chacune de ces phases jouent un rôle crucial dans la performance totale des systèmes de question-réponse. Il peut être évalué de deux manières : intrinsèque (modulaire) et extrinsèque (globale). L'évaluation de NArQAS a été réalisée sur un corpus composé de

questions et textes collectés et extraits à partir Web. Les performances sur l'extraction de la réponse précise montrent un rappel de 68% et une précision de 73%.

1.1 **Présentation de NArQAS**

NArQAS permet de chercher des réponses à partir du Web à des questions factuelles. Nous avons combiné des techniques d'intelligence artificielle, de recherche d'informations, de TALN et de raisonnement automatique pour améliorer les performances de notre système en prenant en compte l'aspect de plusieurs aspects dans les systèmes précédents, notamment en arabe, comme la compréhension automatique des textes, l'intégration de la sémantique et de la logique à la langue arabe, l'utilisation de la technique RTE pour trouver la réponse exacte parmi plusieurs réponses candidates. En effet, l'utilisation de ces deux derniers aspects dans les systèmes de question-réponse a largement été démontrée notamment par des applications en anglais.

Bien que, les outils et les techniques existants aient été essentiellement conçus pour optimiser la performance des technologies traditionnelles de recherche d'informations, nous constatons que leurs performances sont affectées par celles des techniques de traitement automatique de la langue naturelle adoptées. Dans nos travaux, le processus de génération de la réponse précise s'appuie essentiellement sur une étape de représentation logique. Par conséquent, le but de notre système est de répondre aux préoccupations suivantes :

(a) Analyser des questions collectées.
(b) Interroger un moteur de recherche pour chercher le document pertinent
(c) Récupérer des passages de textes contenant les réponses à ces questions.
(d) Effectuer des analyses linguistiques pour la question et leurs passages réponses (analyse morphologique, syntaxique et reconnaissance).
(e) Construire des représentations sémantiques de la question et des passages avec le formalisme du graphe conceptuel.
(f) Déduire des représentations logiques des représentations des graphes conceptuels des questions et des passages de texte.
(g) Appliquer une technique RTE pour trouver la bonne réponse.
(h) Extraire la réponse.

Les tâches **(b)** et **(c)** dépendent sur des techniques de recherche d'informations, les tâches **(a), (d), (e) et (h)** sont des tâches de traitement automatique de la langue naturelle, la

tâche **(g)** est assurée en appliquant les techniques de RTE. Finalement, la tâche **(f)** est résolue à travers des techniques d'intelligence artificielle, en particulier de raisonnement automatique. Notons que le processus qui s'occupe de la représentation logique et de la reconnaissance d'implication textuelle, reste un défi pour la mise en œuvre de tels systèmes en arabe.

1.2 **Outils, ressources et techniques utilisés**

Chaque langue possède ses propres caractéristiques et dispositifs. Ainsi, il semble difficile d'appliquer les mêmes techniques pour toutes les langues. Généralement, la recherche d'une réponse précise à une question en langue naturelle s'appuie principalement sur des techniques de traitement automatique de la langue et de recherche d'informations. Evidemment, les outils de recherche d'informations sont employés plus particulièrement à la recherche de documents et de passages les plus pertinents, tandis que les techniques de traitement de la langue permettent d'améliorer les procédures d'extraction d'informations en offrant la possibilité d'effectuer une analyse approfondie des documents (e.x. la question, des passages, etc). Le choix des outils et ressources utilisés dépend de la fiabilité (le temps de réponse raisonnable), de la couverture (une base de données riche, qui regroupe la totalité des mots arabes) et de l'efficacité (les résultats parvenus sont satisfaisants et répondent aux besoins de l'application). La possibilité de choisir la technique appropriée à chaque type de question atteint des performances proches de la réponse souhaitée tel que l'obtention de la réponse en temps réel.

En effet, l'arabe est une langue très riche. Toutefois, cette richesse nécessite une manipulation particulière, ce qui rend les techniques régulières de traitement de langue naturelle, de recherche d'informations, d'extraction d'informations ou autres, conçues pour d'autres langues, incapables de la manipuler. A ce titre, et malgré les divers efforts, la maturité et l'efficacité de ce type d'outils pour le cas de la langue arabe, est proportionnellement faible par rapport à d'autres langues. Dans notre système, la plupart des modules peuvent bien entendu impliquer sur des techniques et outils externes. Les deux modules centraux (recherche de documents et sélection de passages) reposent sur des outils de recherche d'informations. Les deux autres modules (analyse de la question et extraction de la réponse) reposent sur des modules impliquant de manière plus fondamentale des techniques de traitement de la langue. Le module de représentation logique et de reconnaissance des implications textuelles repose sur des techniques de raisonnement automatique et

d'intelligence artificielle. La suite de cette section, décrit en détails les outils et les techniques intégrés dans le développement de NArQAS.

Outils du traitement automatiques de la langue

Il y a divers outils qui sont utilisés pour le traitement automatique de la langue. Ceux-ci incluent des outils d'analyse morphologique, d'analyse syntaxique, de reconnaissance des entités nommées, etc. Dans nos travaux, les composants de NArQAS s'appuient sur certains de ces outils afin d'effectuer des analyses linguistiques de la question et des passages réponses.

a. Analyseur morphologique : Khoja Stemmer

Dans nos travaux, une analyse morphologique des mots de la question d'entrée et des phrases des passages réponses est effectuée en utilisant Khoja Stemmer [Larkey & Connell, 2001]. Ce dernier a été utilisé dans le cadre d'un système de recherche d'informations développé à l'Université du Massachusetts, aux États-Unis, pour la piste multilingue de TREC-10 en 2001. Cet outil fonctionne en éliminant le suffixe le plus long et le préfixe le plus long, puis associe le mot restant aux motifs verbaux et nominaux pour extraire la racine. Dans leur travail, les auteurs gèrent les lettres faibles (e.g. alif, waw ou yah) et les mots arabes qui n'ont pas de racines. Une implémentation Java de l'algorithme de Shereen Khoja est accessible sur le Web[19].

- https://sourceforge.net/projects/arabicstemmer/

b. Analyseur syntaxique : Stanford parser

Pour chaque question et leurs passages réponse correspondants, nous utilisons l'analyseur syntaxique Stanford [Manning & Jurafsky, 2012]. Ceci est un projet implémenté en Java et développé à l'Université de Stanford. Ce dernier est un outil open source, il prend en charge l'anglais, le chinois, l'allemand et l'arabe. Il est utilisé aussi pour d'autres langues, comme l'italien, le bulgare et le portugais. Dans nos travaux, nous utilisons Stanford parser afin d'identifier les constituants de la question et de leurs passages de texte avec des rôles thématiques et produire les dépendances et les tags des mots.

- https://nlp.stanford.edu/software/lex-parser.shtml

[19] https://sourceforge.net/projects/arabicstemmer/

c. Reconnaissance des entités nommées : ArNER

Il est également très important de souligner qu'une reconnaissance d'entités nommées REN est requise pour presque tous les systèmes de question-réponse qui traitent les questions factuelles. Dans nos travaux, nous utilisons un outil de reconnaissance des entités nommées ArNER [Zribi et al., 2010] qui a été défini dans l'équipe de travail de traitement du langage naturel du Laboratoire MIR@CL. Nous avons choisi ce système car c'est le standard le plus connu parmi quelques systèmes d'analyse réalisés dans notre laboratoire. Plus précisément, pour la question et les passages réponses, ArNER reçoit le texte de la question et des passages pour fournit un fichier XML qui contient toutes les entités nommées de ces deux documents.

- **Outils d'extraction d'informations**

Le processus de la recherche des réponses exige des analyses approfondies des questions ou des passages qui peuvent comporter la réponse exacte. En effet, la reconnaissance des entités nommées peut être considérée comme un outil ou technique d'extraction d'informations. D'ailleurs, le processus d'annotation des entités nommées s'accomplit par le biais d'un jeu d'étiquettes (ou labels) correspondant aux types utilisés pour définir les différents types de la réponse. Dans nos travaux, les entités nommées issues dans la question peuvent jouer un rôle considérable dans l'extraction des réponses potentielles. Plus précisément, après avoir choisi le passage le plus pertinent, nous utilisons les entités nommées de la question et de ce passage pour extraire la réponse précise ou l'entité nommée.

- **Outils de recherche d'informations**

Lors de la phase de recherche de documents, il est possible d'utiliser de nombreux moteurs de recherche comme des outils de recherche d'informations. Dans nos travaux, nous avons utilisé le moteur de recherche google, comme étant une source de données linguistiques, pour extraire les passages pertinents qui sont susceptibles de contenir la réponse précise à une question donnée.

- **Outils d'intelligence artificielle**

Ce processus se fonde également sur des techniques d'intelligence artificielle telles que le raisonnement logique. En effet, la maturité et l'efficacité de ces outils diffèrent selon le niveau de complexité du domaine traité et selon la langue cible. Dans nos travaux, nous

appuyons sur le principe de l'opérateur Φ de [Sowa, 1984] qui associe une formule logique à un graphe conceptuel ou à un vocabulaire.

- **Ressources sémantiques et linguistiques**

Construire un système de question-réponse arabe n'est pas une tâche simple. Pour le faire, nous utilisons un lexique linguistique («المعجم الوسيط»), une ressource sémantique («AWN»), etc. D'ailleurs, nous trouvons peu de ressources sémantiques (par exemple, les thesaurus, les ontologies, etc.) sont disponibles pour l'arabe en comparaison avec les autres langues.

a. WordNet arabe

WordNet arabe (AWN) [Elkateb et al. 2006] est une ressource lexicale pour l'arabe fondée sur le développement de Princeton WordNet pour l'anglais [Fellbaum, 1998]. AWN a une structure d'un thesaurus, il est organisé structuré selon des synsets qui sont un ensemble de synonymes. Ces synonymes sont regroupés afin de décrire le sens (signification) des mots. D'ailleurs, les synsets sont divisés en fonction des parties du discours en quatre types: nom, verbe, adjectif et adverbe. Dans nos travaux, nous utilisons WordNet arabe pour la construction des graphes conceptuels de la question et des passages (spécifiquement dans l'étape d'extraction de concepts). Nous utilisons pareillement WordNet arabe pour la détermination d'implication textuelle entre les représentations logiques de la question et de leurs passages réponses.

b. Dictionnaire «المعجم الوسيط»

Le recours à une ressource linguistique, dans nos travaux, est utile. Plus précisément, nous utilisons le dictionnaire intermédiaire **«المعجم الوسيط»** [Muṣṭafá et al., 2008] qui contient les différentes définitions des mots. En effet, «المعجم الوسيط» est une version du lexique arabe de l'académie de l'arabe à Egypte fournit par la platforme SAFAR[20]. Ce lexique a été utilisé dans deux cas : (i) pour la construction des graphes conceptuels comme ressource de désambiguïsation des mots ambigus ; (ii) ou pour le traitement effectué dans un cas d'erreurs confronté avec les questions commençant par **« من »**. Dans ce cadre, nous utilisons le lexique intermédiaire «المعجم الوسيط» pour tester si le mot qui suit la particule « من » est un verbe.

- **Techniques de RTE**

[20] http://arabic.emi.ac.ma/safar/

Nous discutons maintenant quelles techniques nous avons utilisées dans nos travaux. Nous avons cité dans le chapitre 4 que le problème d'implication textuelle est étudié comme un problème de classification. Pour le faire, plusieurs techniques ont été prises en considération. Ces techniques sont entre autre la mesure de chevauchement de mots, l'apprentissage automatique et la distance sémantique. Plus précisément, pour le chevauchement de mots, nous utilisant la mesure Overlap. Pour la distance sémantique, nous utilisons la mesure Wu-Palmer parmi plusieurs autres mesures de différentes catégories fournies l'API AWN (p.ex. edge counting, Wu Palmer, Li, etc). Finalement, pour l'apprentissage, nous utilisons, un classificateur d'arbre de décision J48 de WEKA [Witten & Frank, 1999].

2. **Conception et implémentation de NArQAS**

Pour concevoir notre système NArQAS, nous avons adopté l'architecture généralement utilisée pour un système de question-réponse. Notre système se situe en aval des modules d'analyse de la question et du texte constitué des passages répondant à cette question. Tout d'abord, le processus de recherche débute par l'analyse de la question posée jusqu'à atteindre la réponse précise. Néanmoins, si les éléments de la question ne sont pas identifiés correctement, il reste peu de chances de trouver la réponse. La plupart des systèmes de question-réponse reposent sur une architecture classiquement fondée sur trois ou quatre modules. Ces modules s'appuient principalement sur des techniques de traitement automatique de la langue aussi des techniques de recherche d'informations. Plus particulièrement, les outils de recherche d'informations servent à la recherche des documents et des passages les plus pertinents, tandis que les techniques de traitement de la langue permettent d'améliorer les procédures d'extraction d'informations en offrant la possibilité d'effectuer une analyse approfondie de la question et des documents. Dans nos travaux, nous ajoutons d'autres modules qui importent pour la sélection de la réponse et pour lesquels des traitements d'analyse sémantique, de raisonnement automatique et de RTE sont réalisés.

2.1 **Architecture de NArQAS**

NArQAS avait simplement été mentionné dans la section précédente comme étant un outil permettant d'obtenir des réponses précises à des questions en arabe, nous détaillons ici son architecture, puis nous terminons par montrer le déroulement de ses principaux composants ainsi que son fonctionnement. L'architecture schématisant le fonctionnement de notre approche est illustrée par la figure 5.29. La conception du système NArQAS emploie

généralement une architecture pipeline qui assemble six modules principaux à savoir: l'analyse de la question, la récupération des documents, l'extraction des passages, l'analyse des passages, la représentation logique et l'extraction de la réponse. Chaque module repose sur des techniques et traitements particuliers. Par exemple, l'analyse des questions repose sur des techniques liées au traitement automatique des langues, la recherche de documents repose sur des techniques de la recherche d'informations afin d'obtenir les documents pertinents par rapport à la question ainsi qu'au domaine de l'extraction d'informations pour extraire la réponse précise attendue, etc. Chacun de ces composants mérite d'être évalué intrinsèquement, ou leur assemblage est évalué dans son ensemble. Nous dressons dans la suite de cette section une description des principes des différents modules qui composent cette architecture.

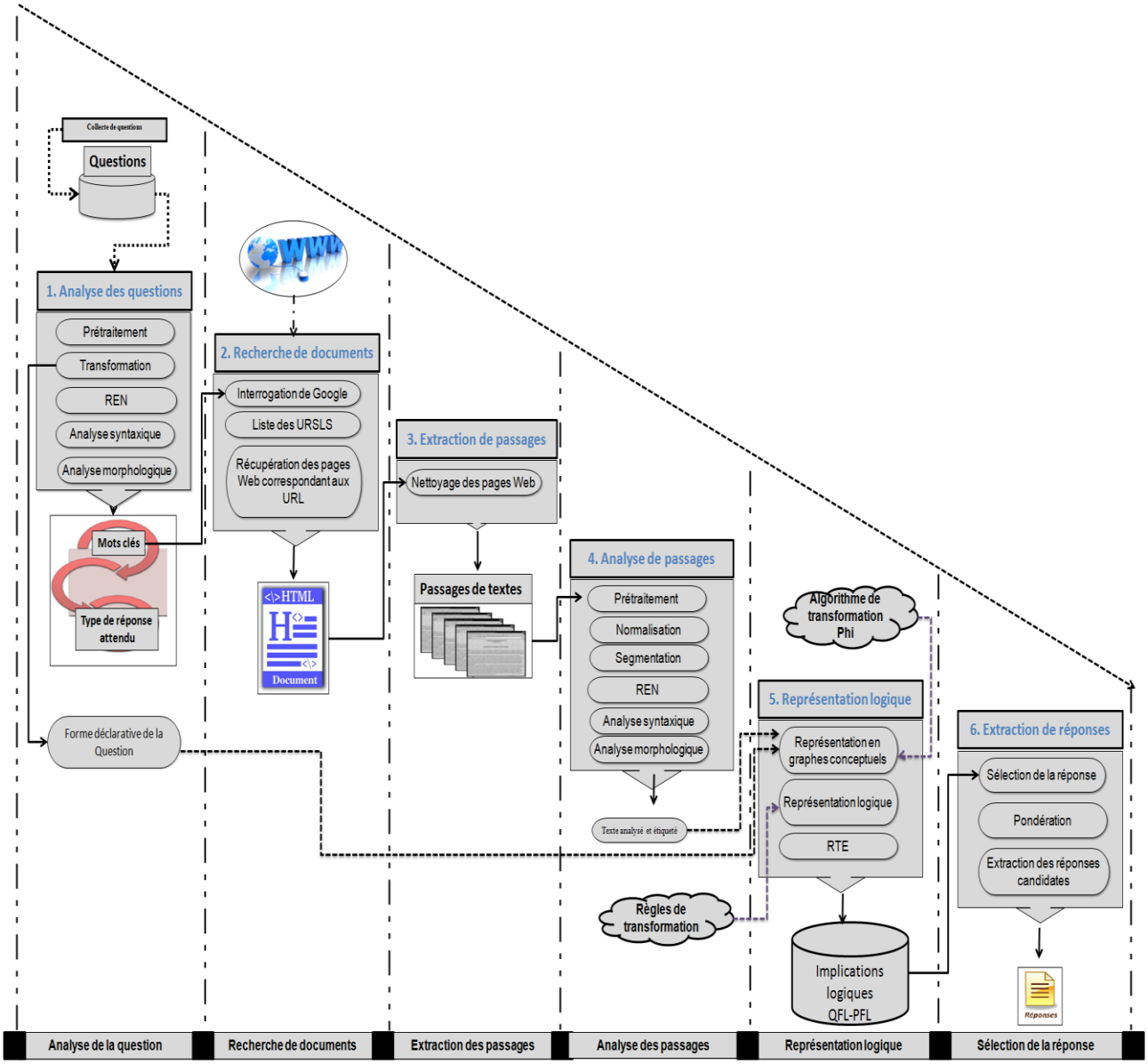


Figure 5.29: Architecture de NArQAS

La figure 5.29 montre la façon dont chaque composant se rapporte à l'autre. Nous détaillerons par la suite les différents modules intervenant dans la chaîne de traitement, soit de l'analyse de la question jusqu'à l'élaboration de la réponse exacte. Notons que notre système proposé est doté d'une architecture complexe et s'appuie sur des techniques de recherche plus élaborées à savoir le raisonnement logique et RTE. La conception de ce système a largement contribué au développement des systèmes de question-réponse, notamment pour l'arabe.

2.2 **Vue de fonctionnement**

Pour décrire le fonctionnement des composants et trouver la réponse exacte à cette question, nous suivons le schéma de la figure 5.30 qui illustre la vision des opérations successives mises en œuvre dans le système développée tout au long de cette thèse. Par conséquent, chaque opération peut contenir plusieurs actes séquentiels.

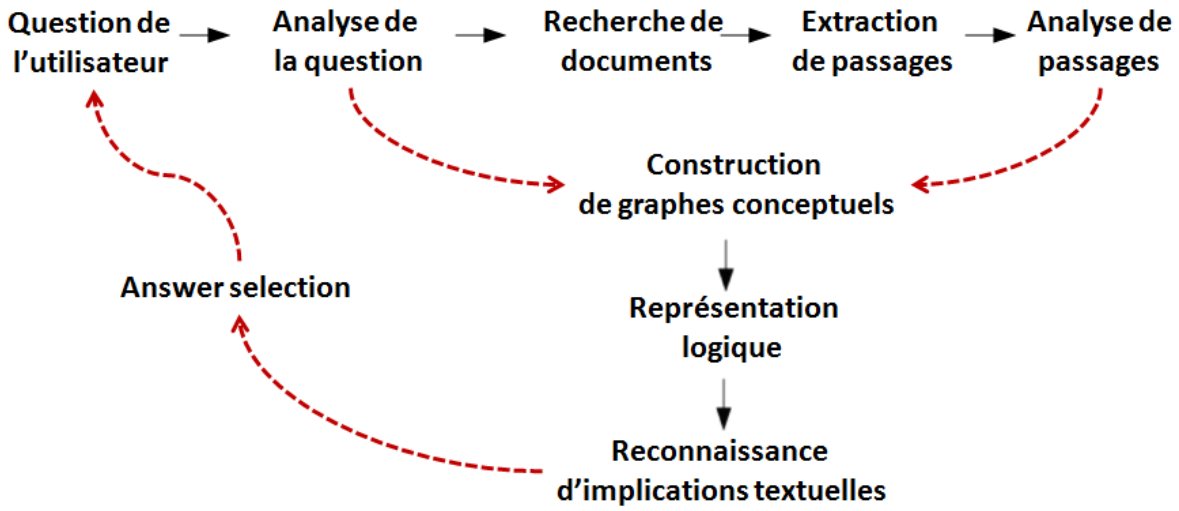


Figure 5.30: Schéma de fonctionnement de notre système

De façon schématique, le processus de la recherche d'une réponse précise à une question factuelle comprend les modules suivants :

1. L'utilisateur écrit en langue arabe sa question.
2. Le module d'analyse de la question se charge d'identifier tous les mots clés de la question, son type et son type de réponse attendu en générant la forme déclarative correspondante à chaque question et appliquant une liste d'analyse linguistique (syntaxique, morphologique, etc.).

3. Le module de recherche de documents utilise le moteur de recherche Google qui accepte les mots-clés en entrée et produit des documents qui sont étroitement liées à ces mots-clés. Dans ce cadre, la recherche de documents repose sur un appariement entre les termes de la question et ceux des documents. La recherche ne se limite pas à trouver des ressources référencées par des mots clés, mais tente d'identifier des passages pertinents qui contiennent des réponses aux questions.
4. Le module d'extraction de passages identifie parmi les documents retenus les passages susceptibles de contenir les réponses exactes aux questions posées.
5. Le module d'analyse de passages se charge de segmenter les passages en phrases et étiquette les entités nommées qui s'y trouvent et applique une liste d'analyse linguistique (syntaxique, morphologique, etc.).
6. Le module de construction de graphes conceptuels consiste à représenter le document (passages de texte et question) dans des graphes conceptuels pour faciliter leur transformation en des formes logiques (ici, structures d'arguments-prédicats).
7. Le module de représentation logique pend en considération la transformation des graphes conceptuels des passages de texte ou de la question en logique de premier ordre.
8. Le module d'implication textuelle vient de déterminer les relations d'implication textuelle entre les représentations logiques des passages de texte et de la question.
9. Enfin, le module d'extraction de la réponse se charge de sélectionner la réponse la plus cohérente et exacte parmi d'autres candidates (ici, le passage pertinent).

2.3 **Composants de NArQAS : entrées-sorties**

Nous décrivons, dans cette section, les différents composants de NArQAS. En effet, chaque module est considéré comme une boîte : nous avons une entrée, un traitement et une sortie. D'ailleurs, la structure basée sur les modules d'un système de question-réponse est accompagnée d'une supposition sur la composition de ses composants. Il est facile de constater que les erreurs en cascade que le processus de question-réponse se déplace à travers les modules en aval. Par conséquent, l'amélioration de performances des modules individuels affaiblit l'erreur à chaque étape du traitement qui, à son tour, devrait permettre de maximiser

la précision globale du système. L'architecture du système NArQAS est composée des composants suivants :

- ***Analyseur des questions***

  - ✓ But:

Le but primordial du module « analyseur des question » est d'obtenir les caractéristiques pertinentes de la question qui seront utiles dans les étapes suivantes. Ce module implique plusieurs sous étapes comme le prétraitement, la transformation, l'identification des entités nommées, l'analyse morphologique, l'analyse syntaxique, etc. Ce module, qui est un composant de base de tout système question-réponse. Notamment, il est essentiel qu'une question soit analysée aussi finement que possible pour qu'une réponse correcte lui soit apportée. De ce fait, le système n'arrivera pas à trouver de réponse si l'analyse de la question est erronée.

  - ✓ Entrée:

Une question en langue arabe (question factuelle de type: (What, Where, When, Who, How) (ما، أين، متى، من كم)).

  - ✓ Sortie:

L'analyse des questions produit le type de la réponse attendu, les mots-clés et la forme déclarative, les entités nommées, l'analyse de dépendance, POS, la tige, etc.

- ***Extracteur de passages***

  - ✓ But:

Ce module se charge d'extraire les passages candidats susceptibles de contenir la réponse. Cette étape se révèle particulièrement capitale et complémentaire à l'analyse de la question et la recherche de documents pour trouver la réponse précise car les systèmes de question-réponse ne peuvent trouver une réponse à une question que si elle est présente dans les documents sélectionnés. Les meilleurs passages sélectionnés sont des textes ou les phrases correspondant à la réponse recherchée. La stratégie pour le faire consiste le plus souvent à extraire les passages ou les phrases comportant au moins un mot de la question ou une entité du même type que la réponse attendue. Généralement, le passage candidat est composé d'un

bloc d'une, de deux ou trois phrases regroupant la phrase réponse complétée par la phrase précédente et la phrase suivante.

- ✓ Entrée:

Des liens des pages web des documents retenus par l'étape de recherche de documents.

- ✓ Sortie:

La sortie correspond à un fichier texte contenant les passages répondant aux questions.

- ***Analyseur de passages***

- ✓ But:

Ce composant se charge d'analyser les passages candidats susceptibles de contenir la réponse. Répondre à des questions précises requiert une analyse plus en profondeur des passages afin d'en extraire l'information pertinente. En effet, il ne suffit pas de chercher des passages contenant les mots de la réponse à l'identique pour trouver une réponse correcte. Il faut approfondir leur analyse afin d'extraire la réponse. Cet approfondissement assure en générale un enrichissement de chaque passage candidat. Parmi les enrichissements les plus fréquents, des prétraitements et normalisations de textes, des segmentations en phrases, la détermination des entités nommées sont prises en compte. Enfin, les passages ainsi segmentés sont annotés avec des étiquettes morphologiques (Tige) et analysés avec Stanford pour obtenir des dépendances syntaxiques.

- ✓ Entrée:

Des passages de textes.

- ✓ Sortie :

Passages de textes segmentés, analysés et annotés

- ***Générateur des graphes conceptuels***

- ✓ But:

Dans ce module, nous proposons de représenter la question et les passages réponses en des graphes conceptuels. Particulièrement, la sémantique de la question et des textes se transforme en sémantique de modèles conceptuels à un haut niveau d'abstraction, en termes de concepts et de relations. Ce module prend en considération les résultats des analyses

effectuées précédemment. Il produit ainsi pour chaque question et chaque phrase du passage un graphe conceptuel.

- ✓ Entrée:

Les résultats des analyses effectuées sur les passages et les questions

- ✓ Sortie:

Graphe conceptuel pour chaque phrase de passage et pour chaque question.

- ***Générateur de formes logiques***

- ✓ But:

Dans ce module, les graphes conceptuels des questions et les phrases des passages sont traduits en des représentations logiques qui sont illustrés par une structure de Prédicat-arguments. En effet, la transformation est assurée par un algorithme de transformation. Ce module reçoit donc en entrée le graphe conceptuel pour chaque question et passage et produit la représentation logique correspondante.

- ✓ Entrée:

Graphe conceptuel pour la question et phrase de passage.

- ✓ Sortie :

Le résultat est un ensemble de formules logiques associées aux questions et phrases de passages réponses.

- ***Détecteur d'implication textuelle***

- ✓ But:

Ce module a pour objectif de détecter par l'intermédiaire des métriques de RTE de trouver les passages qui on une implication avec la question. Ce module tient compte une série de caractéristiques extraite pour déterminer la tâche de l'implication textuelle entre chaque paire de forme logique d'un passage et d'une question. Ce module utilisé trois caractéristiques pour accomplir cette tâche et renvoie en sorties les passages ayant une relation d'implication avec la question.

- ✓ Entrée:

Des représentations logiques des questions et de passages susceptibles de répondre à cette question.

- ✓ Sortie :

La liste de passages retenus avec implication avec leurs scores de confiances.

- ***Extracteur de la réponse***

✓ But:

Enfin, le dernier module consiste à extraire des réponses candidates de ces passages en s'appuyant sur certaines caractéristiques déduites de la question et la façon dont elles se retrouvent au niveau des passages pour détecter une réponse dans une phrase candidate, comme le type de la réponse attendu. En effet, ce module constitue le dernier maillon de la chaîne de traitement de notre système de question-réponse proposé. Ce module prend en entrée des passages appelés candidats issues de détermination d'implication textuelle. Enfin, nous sélectionnons l'entité nommée du type attendu la plus proche des mots de la question.

✓ Entrée:

Des passages ayant implication avec la question ainsi que leurs scores de confiances.

✓ Sortie :

La réponse produite est présentée à l'usager (un passage et l'entité nommée)

2.4 **Détails de l'implémentation**

L'étape de conception de l'architecture se termine lorsqu'une description structurelle du système désiré est réalisée et qu'un ensemble de types de composants primitifs sont représentés en un langage de programmation. Dans cette section nous présentons quelques détails de l'implémentation de notre système NArQAS. En effet, nous avons développé une interface en Java gérant en amont tous les traitements nécessaires aux tâches demandées par l'utilisateur afin d'obtenir une réponse précise à une question en arabe. Cette interface comprend les traitements suivants tels que : l'analyse de la question, la recherche de documents, la sélection des passages pertinents, l'analyse des passages, la représentation logique des questions et passages et la détection de l'implication textuelle entre eux ; et enfin, l'extraction de la réponse précise. Le langage de programmation que nous avons choisi pour l'implémentation est JAVA.

En effet, la mise en œuvre de l'approche proposée dans le cadre d'un système enrichi par souci d'une meilleure performance de la question-réponse arabe. En effet, NArQAS peut être décrit schématiquement comme un enchaînement séquentiel de tâches caractéristiques, où chacune contribue à produire l'entrée de l'étape suivante. Dans la suite de cette section, nous avons décrit les principaux modules implémentés collaborant dans une architecture fortement

modulaire pour permettre l'extraction de la réponse précise. En effet, chacun de ces modules joue un rôle crucial dans la performance globale de NArQAS.

- **Prétraitement et analyse des questions**

Le système se présente donc sous la forme d'une simple invite à taper une question. En effet, le module d'analyse de la question suppose que chaque question doive être une phrase déclarative simple qui est composée d'une séquence de mots. Puis, l'utilisateur clique sur le bouton **« Transform »**, le système extrait les mots clés de cette question, le type de la réponse attendue comme une preuve utile pour extraire la réponse précise. En pratique, le type de question est généralement lié au type de la réponse attendue, qui à son tour est généralement lié aux types disponibles pour les entités nommées. L'idée est que les questions factuelles se répartissent en plusieurs types distinctifs, tels que «lieu», «organisation», «personne», etc. Ensuite, une forme déclarative pour chaque question est générée. En outre, l'utilisateur a la possibilité de poser de nouvelles questions en cliquant sur le bouton **« Clear »**. Enfin, l'analyse linguistique de la question fournit son analyse morphologique, syntaxique et la reconnaissance des entités nommées issues dans cette section. Dans ce cadre, l'utilisateur clique sur le bouton correspondant concernant l'analyse, le système fournit le résultat correspondant à cette action. Pour donner une compréhension plus profonde des divers composants et valider les phases précédentes, des exemples courants sont illustrés et la question suivante Q-teste « أين يوجد مقر منظمة اليونسكو ؟ » servira de fil conducteur suivant :

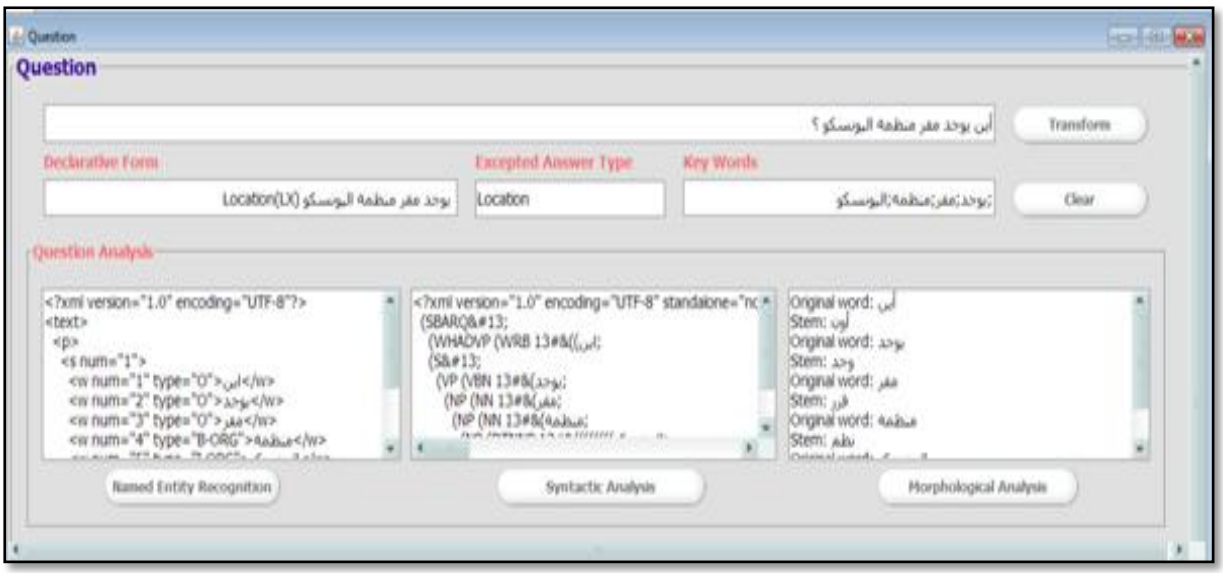


Figure 5.31: Prétraitement et analyse de la question Q-teste

- **Recherche et extraction des passages pertinents**

Nous avons utilisé le moteur de recherche google pour trouver les documents contenant au moins une réponse aux questions. A cet égard, l'interrogation d'un moteur de recherche peut accélérer la récupération de documents en ligne, mais nécessite un traitement hors ligne de ces documents. Précisément, au lieu de renvoyer un certain nombre de documents tels que google, notre module de récupération de passage renvoie des passages contenant des phrases de différentes longueurs qui peuvent être des réponses aux questions des usagers. La figure 5.32 montre un fonctionnement réel de notre module de récupération de passage. Quand l'utilisateur clique sur le bouton "Générer HTML", le système récupère une page à partir de son URL. La dernière étape consiste à transformer chaque page Web obtenue dans un format ".txt". Les textes étant en format ".html", et étant donné que l'application envisagée est la modélisation statistique du langage, il semble justifié de les mettre dans le format ".txt". Pour cela, on enlève toutes les balises HTML pour chaque page récupérée. Il est possible, soit est de garder le texte de son propre travail de construction de corpus, ou de l'ignorer.

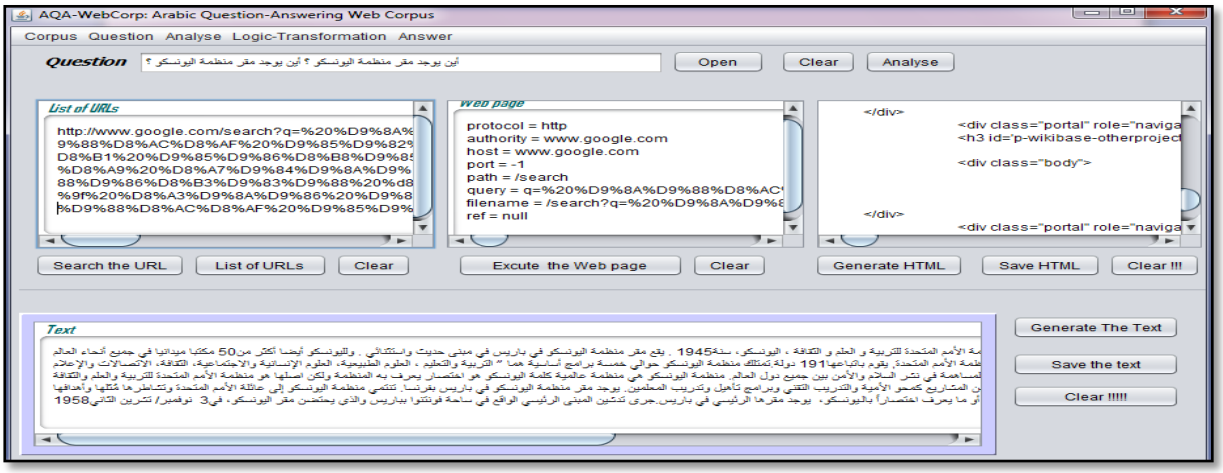


Figure 5.32: Recherche et extraction des passages répondant à la question Q-teste

A partir de la question citée précédemment, considérons l'extrait des passages pertinents suivant susceptibles de contenir la réponse précise à la question **Q-teste** :

تأسست منظمة الأمم المتحدة للتربية و العلم و الثقافة، اليونسكو، سنة 1945. يقع مقر منظمة اليونسكو في باريس في مبنى حديث واستثنائي . ولليونسكو أيضا أكثر من 50 مكتبا ميدانيا في جميع أنحاء العالم
يوجد مقر منظمة اليونسكو بباريس فهناك المقر الرئيسي لمنظمة اليونسكو. اليونسكو هي وكالة متخصصة تتبع منظمة الأمم المتحدة, يقوم باتباعها191 دولة.تمتلك منظمة اليونسكو حوالي خمسة برامج أساسية هما " التربية والتعليم ، العلوم الطبيعية، العلوم الإنسانية والاجتماعية، الثقافة، الاتصالات والإعلام ".
الهدف الرئيسي من انشاء منظمة اليونسكو هو المساهمة في نشر السلام والأمن بين جميع دول العالم. منظمة اليونسكو هي منظمة عالمية كلمة اليونسكو هو اختصار يعرف به المنظمة ولكن اصلها هو منظمة الأمم المتحدة للتربية والعلم والثقافة
تدعم اليونسكو العديد من المشاريع كمحو الأمية والتدريب التقني وبرامج تأهيل وتدريب المعلمين. يوجد مقر منظمة اليونسكو في باريس بفرنسا. تنتمي منظمة اليونسكو إلى عائلة الأمم المتحدة وتشاطر ها مُثلها وأهدافها
منظمة الأمم المتحدة للتربية و العلم والثقافة أو ما يعرف اختصاراً باليونسكو، يوجد مقر ها الرئيسي في باريس.جرى تدشين المبنى الرئيسي الواقع في ساحة فونتنوا بباريس والذي يحتضن مقر اليونسكو، في 3 نوفمبر/ تشرين الثاني 1958

Figure 5.33: Passages de texte extraits du Web pour la question Q-teste

Post traitement et analyse des passages

Afin d'augmenter la chance de trouver la réponse précise, nous avons décidé d'effectuer l'analyse des passages de texte qui contiennent cette réponse. D'ailleurs, chaque passage candidat est approximativement composé d'un bloc de deux ou trois phrases regroupant la phrase réponse complétée par la phrase précédente et la phrase suivante. De ce fait, des traitements linguistiques sont mis en œuvre par des techniques de TALN, d'EI ou d'autres permet de normaliser les mots ainsi d'extraire les entités nommées de ces passages sélectionnés. Ce traitement accepte en entrée un texte arabe en format «.txt » et génère un texte annoté et analysé. Ce traitement se compose de plusieurs sous-étapes séquentielles. Tout d'abord, une étape de segmentation détermine la division du texte en jetons (phrases). Puis, le prétraitement est effectué pour éliminer les mots vides. Ensuite, les passages réponses parfois contiennent des mots étrangers, des caractères spéciaux, des nombres (e.g. ',:,?, \, $,…), etc. Enfin, le module d'analyse des passages tient compte la reconnaissance d'entités nommées pour déterminer l'ensemble des entités nommée dans ces passages ; extrait l'analyse syntaxique de ces passages et effectue l'analyse morphologique. Ce module délivre comme résultat final un fichier XML à chaque étape d'analyse et à l'étape de normalisation pour construire notre corpus de questions-passages.

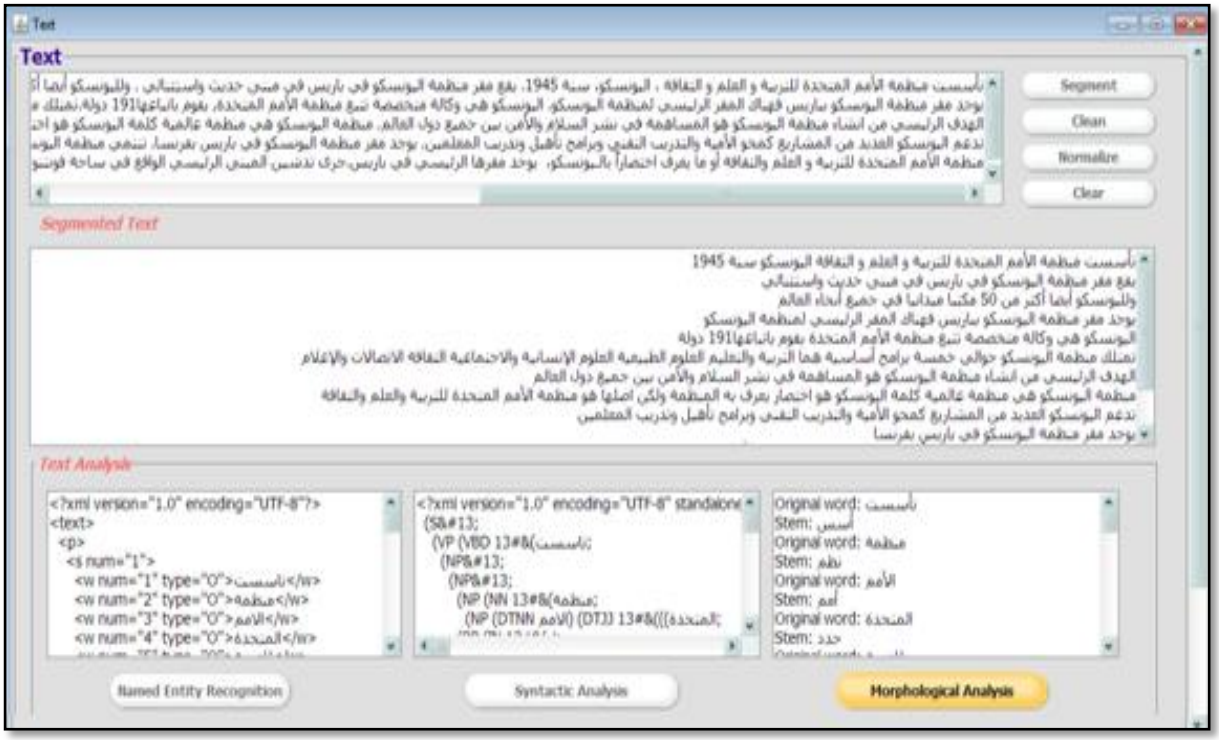


Figure 5.34: Post traitements et analyses des passages

Graphes conceptuels et représentations logiques

Après avoir analysé la question, extrait leurs passages correspondants à partir du web et les analysés, nous avons construit leurs graphes conceptuels afin de les transformer en des représentations logiques. De ce fait, la construction des graphes prennent en considération les résultats des modules d'analyse de la question ou de leurs passages. D'ailleurs, ce module a construit un graphe pour chaque phrase. Enfin, notre module transforme ces graphes conceptuels en des représentations logiques correspondantes. Les graphes conceptuels et les représentations logiques de la question et de leurs passages sont illustrés respectivement dans les figures 5.35 et 5.36.

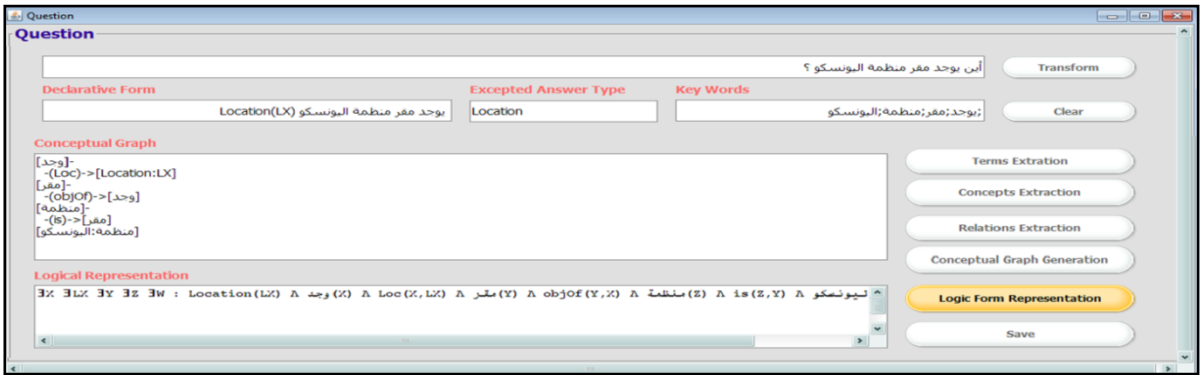


Figure 5.35: Graphe conceptuel et représentation logique de la question Q-teste

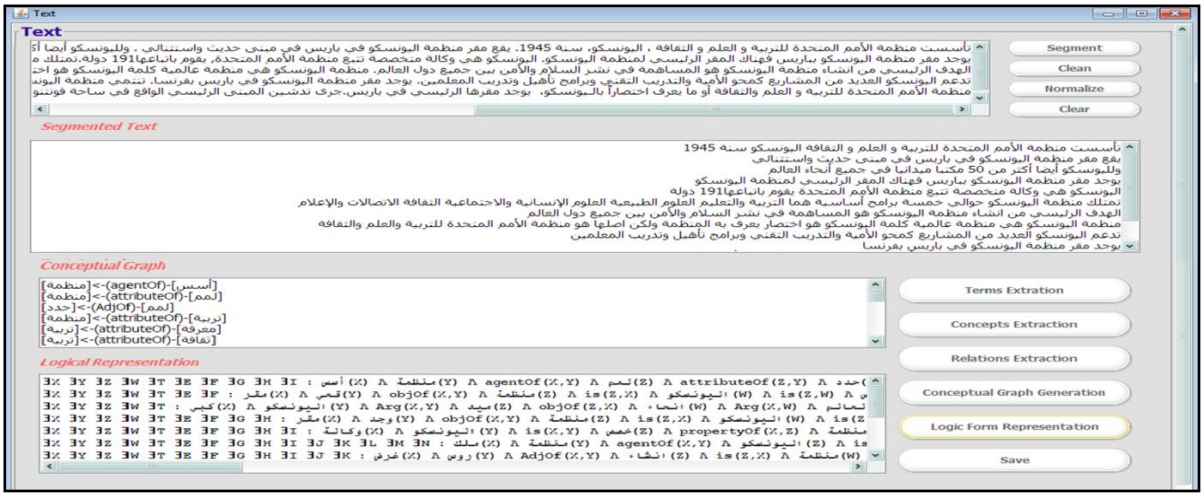


Figure 5.36: Graphes conceptuels et représentations logiques des passages de la question Q-teste

Implication textuelle et extraction de la réponse

A ce stade, le module d'implication textuelle et d'extraction de la réponse est mis en œuvre. D'abord, pour déterminer l'implication textuelle, notre système tient compte du résultat de l'étape de la représentation logique de la question et de leurs passages

correspondants qui est sauvegardé dans un fichier XML. Ensuite, pour extraire la réponse précise, NArQAS tient compte du résultat de l'implication textuelle. Le traitement débute par déterminer la représentation logique de la question et de leurs passages correspondants puis détermine l'implication textuelle entre ces représentations. Bien évidemment, lorsque l'utilisateur pose sa question, le système a pour but de rendre accessible la réponse qui lui satisfait. De ce fait, parmi un nombre de passages réponses candidats retenus par l'étape d'implication textuelle, NArQAS choisit celui le plus pertinent en fonction de son score de confiance. Enfin, après la sélection du passage réponse, la réponse précise sera affichée dans un champ de texte (figure 5.37).

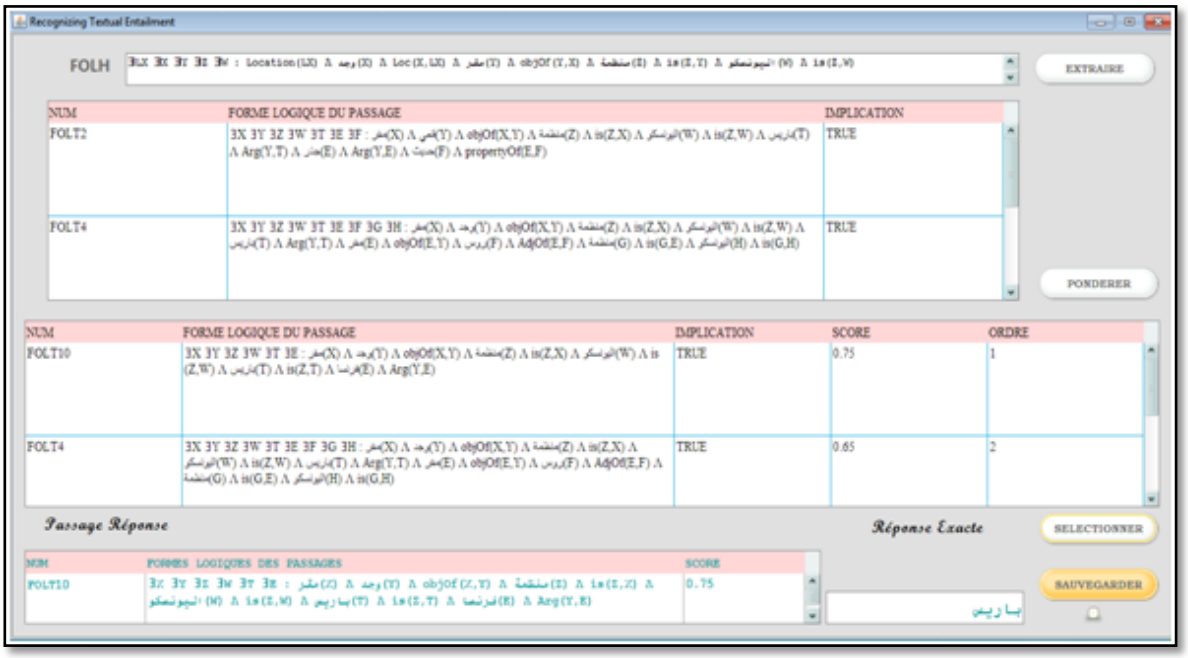


Figure 5.37: Implication textuelle et extraction de la réponse précise à la question Q-teste

3. Evaluation et résultats obtenus

L'évaluation est une étape essentielle dans le développement d'une application informatique pour le TALN, et en particulier pour un système de question-réponse. L'évaluation d'un système de question-réponse peut être faite pour l'ensemble du système et / ou pour chaque module, en particulier le module d'extraction de passages. Elle étant un travail indispensable pour permettre des avancées dans ce domaine de recherche expérimental. En effet, évaluer un système de question-réponse est une problématique car il est difficile de définir ce qu'est une bonne réponse. De ce fait, évaluer d'une façon pointue permet de visualiser d'une façon précise ce qui n'est pas correct, mais aussi d'évaluer la rentabilité de certains traitements. Dans ce cadre, quand un système propose une réponse, il

faudrait qu'il évalue si les passages candidats issus de l'étape de détermination d'implication textuelle sont corrects ou non, et pour cela il se fonde sur une évaluation de la correspondance entre le passage candidat et la question.

3.1 **Ensemble de données pour l'évaluation**

Comme nous avons déjà précisé dans le troisième chapitre, pour évaluer notre système, un corpus (passages de texte et questions) en arabe a été construit. Nous avons exécuté plusieurs types de traitements sur les passages et les questions. En effet, pour construire ce corpus, nous avons effectué une recherche sur le web à l'aide du moteur de recherche Google. D'ailleurs, pour évaluer la qualité de notre système de question-réponse arabe, il faut un nombre important de questions. La première difficulté est d'être en mesure d'élaborer un corpus réaliste de questions, posées par des utilisateurs réels, et correspondant aux besoins réels. Il faut donc des données où sont fournis les questions et les textes susceptibles de contenir des réponses. En effet, notre corpus est construit par des données réelles. Nous avons recueilli 250 questions, soit 25 questions traduites du TREC, 25 questions traduites du CLEF, 100 questions issues des forums et 100 des FAQ. Le choix de ces questions a été guidé par des requêtes les plus fréquemment posées dans les sites web. D'ailleurs, la collection des questions couvrent cinq domaines différents, tels que (**التاريخ ; العالم ; صحة و طب; رياضة; ثقافةو إكتشافات; والإسلام**). Dans ce cadre, les questions sont de types « questions factuelles » dont la réponse attendue est une entité nommée. En l'occurrence, les questions et les passages sont présentés en format TXT.

3.2 **Mesures d'évaluation utilisées**

En pratique, l'évaluation de l'efficacité d'un système de question-réponse se fait par une métrique (point de vue comparatif et quantitatif) ou au niveau suffisance utilisateur (point de vue qualitatif et applicatif). En fait, il y en a beaucoup de métriques d'évaluation sont utilisées. Cependant, couvrant toutes ces mesures n'est pas dans le champ d'application de nos travaux de thèse. Seules les métriques les plus utilisées sont mentionnées dans cette section. Ces mesures sont les mesures de recherche d'informations classiques, le rappel, la précision et la F-mesure.

(a) La Précision :

La précision est la proportion de réponses correctes parmi les oui trouvées par NArQAS.

$$\text{Précision} = \frac{\text{Nombre de réponses «oui positives» retournées}}{\text{Nombre de réponses "oui"retournées}}$$

(b) Le Rappel :

Le rappel est la proportion de nombre de réponses oui correctes retournées par NArQAS par rapport au nombre de réponses oui attendues

$$\text{Rappel} = \frac{\text{Nombre de réponses «oui positives» retournées}}{\text{Total de réponses oui attendues}}$$

La Précision et le Rappel peuvent être à la fois combinés dans une moyenne harmonique pondérée appelée F-mesure.

$$\text{F} - \text{mesure} = \frac{2 * \text{Precision} * \text{Rappel}}{\text{Precision} + \text{Rappel}}$$

3.3 **Résultats obtenus par NArQAS**

Dans cette section, nous présentons successivement les résultats obtenus par NArQAS. En fait, l'évaluation de NArQAS repose généralement sur la validité d'une réponse individuelle supportée par un passage candidat. Dans nos travaux, après la tâche de classification des implications, nous notons que NArQAS produit pour chaque question un nombre de réponses qui sont soient « oui » positives, soient « oui » négatives. En effet, une réponse « oui » positive est un passage retenu avec implication « oui » avec la question et qui contient la réponse à cette question. En revanche, une réponse « oui » négative est un passage retenu avec implication « oui » mais qui ne répond pas à la question. Pour évaluer notre système, seules les réponses « oui » positives sont comptabilisées en utilisant les trois métriques : la précision, le rappel et la F-mesure.

Tableau 5.27: Résultats des expériences menées par NArQAS

Ensemble de questions	Nombre de passages	Réponses Oui retournées	Réponses Oui attendues	Réponses *Oui « + » retournées*	*Précision*	Rappel	F-mesure
Expérience 1: 50 questions	500	189	213	147	0,78	0,69	0,73
Expérience 2: 115 questions	1150	650	714	481	0,74	0,67	0,71
Expérience 3: 250 questions	2500	1690	1720	1143	0,68	0,66	0,67
Moyenne					**0,73**	**0,68**	**0,70**

Pour le faire, nous amenons une série de trois expérimentations sur notre corpus. A la première expérience, nous évaluons NArQAS avec 50 questions. A chaque question, nous prenons en considération les passages correspondants. Notons que notre module d'extraction de passages détermine pour chaque question entre 9 et 14 passages. Dans l'étape d'évaluation, nous prenons pour chaque question les 10 premiers passages. A la deuxième expérience, nous amenons 115 questions. Enfin, nous avons accompli 250 questions et 2500 passages pour la troisième expérience. A titre d'exemple, sur 50 questions NArQAS a proposé 189 réponses (passages) « oui » dont 147 sont des réponses correctes (oui positives) parmi 213 attendues. Le tableau 5.27 présente un exemple montrant les résultats de ces expériences.

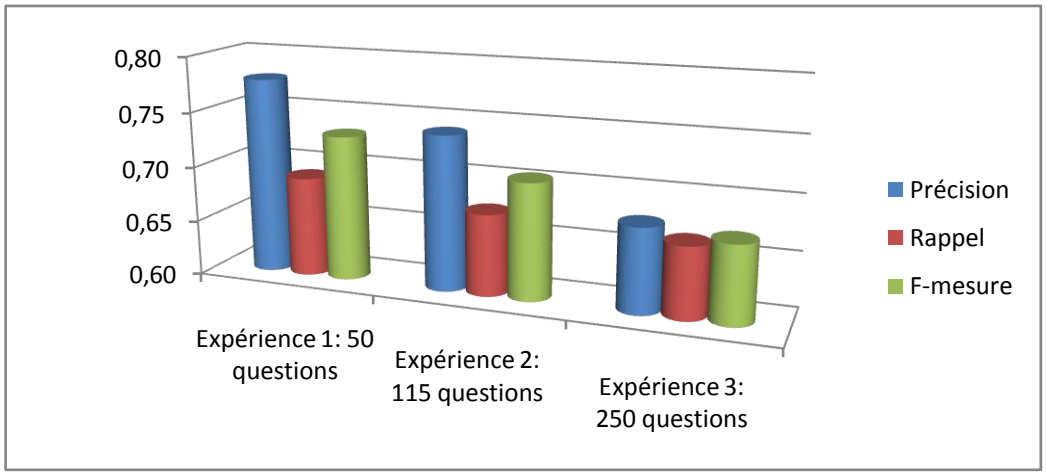


Figure 5.38: Performance de NArQAS

Les résultats de notre système, donnés par le tableau 5.27, montrent une précision et un rappel successivement égaux à 73% et 68% en moyenne, ce qui constitue un bon niveau pour ce type de tâche.

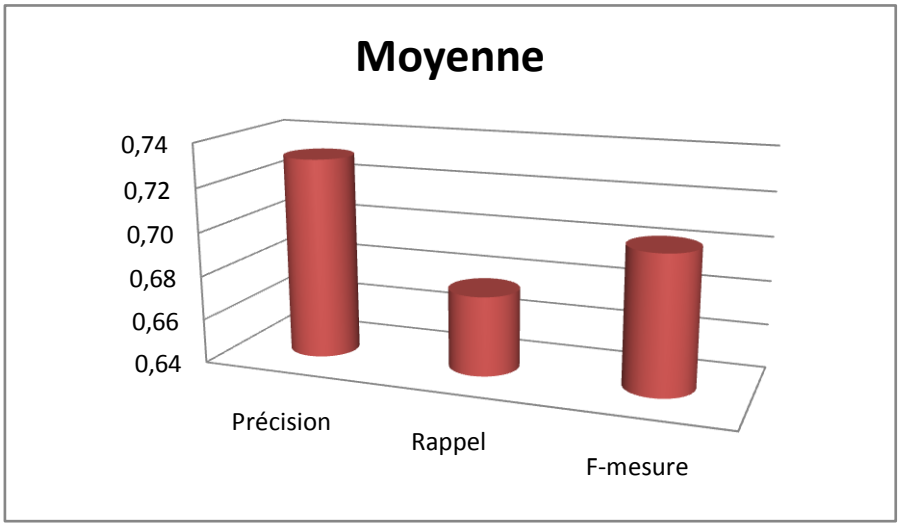


Figure 5.39: Performance de la moyenne

4. Analyse des résultats expérimentaux

Les résultats obtenus par NArQAS sont encourageants. Ces résultats témoignent d'une part l'efficacité de notre système, d'autre part prouvent la performance de la l'approche proposée pour la langue arabe. Ces résultats, montrent également l'importance de la représentation sémantique et logique de la question et des passages ainsi de la reconnaissance d'implication textuelle à la question-réponse en arabe. D'ailleurs, les résultats obtenus lors de nos expériences s'expliquent par les décisions prises lors de la préparation des données et de la réalisation de l'approche proposée. Nous avons réalisé les étapes de l'approche proposée dans le but d'implémenter un système complet de réponse à une question en langue arabe. Dans ce cadre, les résultats de chaque module fournissent certains problèmes. Nous pouvons tenter d'améliorer chaque étape dans le processus de recherche de la réponse en solutionnant les cas d'erreurs que nous avons confronté. Ces erreurs et problèmes rencontrés sont présentés dans la suite de cette section.

4.1 Cas d'erreurs et traitements d'amélioration

Nous présentons, dans cette section, quelques erreurs et problèmes rencontrés lors de la conception et du développement de NArQAS. Nous présentons également les pistes d'améliorations possibles. D'ailleurs, le fait d'améliorer ces résultats montre l'apport de notre approche en comparaison avec les systèmes de question-réponse arabes. En effet, nous avons mentionné précédemment que l'architecture de notre système NArQAS se compose de plusieurs composants. Chacun de ces composants peut affecter les résultats globaux de NArQAS. Dans nos travaux, parmi les cas d'erreurs que nous avons confrontés, celles proviennent de l'analyseur syntaxique Stanford. Par exemple, dans le cas d'une question débutée par « **من** », si un verbe singulier masculin assiste immédiatement après « **من** », il est détecté comme un nom, par Stanford. Par conséquent, les relations qui sont extraites afin de construire les graphes ainsi que les graphes eux-mêmes sont construits d'une façon erronée. Pour résoudre ce problème, nous testons le mon avec le dictionnaire " Intermediate Lexicon (**المعجم الوسيط**)" à l'aide de la plate-forme SAFAR. Si ce mot est un verbe, nous retournons le pattern correspondant.

D'autres cas d'erreurs sont dus à l'application des règles de [Abouenour, 2014]. En effet, l'extraction des règles, dans plusieurs cas, sont effectuées en appliquant une technique basée sur des règles proposées par Abouenour, celle-ci est similaire à celle de [Hensman & Dunnion, 2004] et adaptée à la langue arabe. En revanche, dans nos travaux, nous confrontons

des cas spécifiques où ces règles ne sont pas suffisantes. Pour cela, nous proposons de nouvelles règles pour les résoudre et les améliorer.

- ***Règle 1 :***

Soit la phrase Phr1 suivante : "**الاتحاد الأوروبي هو جمعية دولية للدول الأوروبية يضم 28 دولة**", en appliquant les règles de [Abouenour, 2014], nous obtenons les relations entre les concepts affichées dans la figure 5.40:

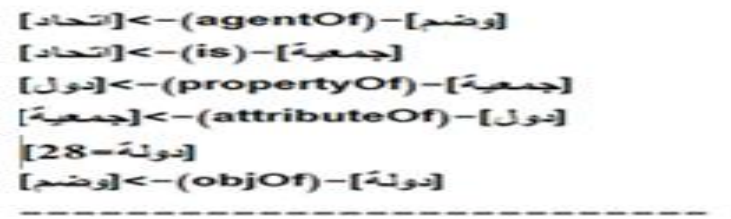


Figure 5.40: Relations retenues en appliquant les règles de [Abouenour, 2014]

Nous constatons qu'il ya une perte d'informations importantes tels que ((**الاتحاد**, **الأوروبي**), (**الدول الأوروبية**)). La dépendance qui relie les mots « **الاتحاد**, **الأوروبي**» et «**الدول**, **الأوروبية**» obtenue par Stanford est « dep », comme le montre la figure 5.41 :

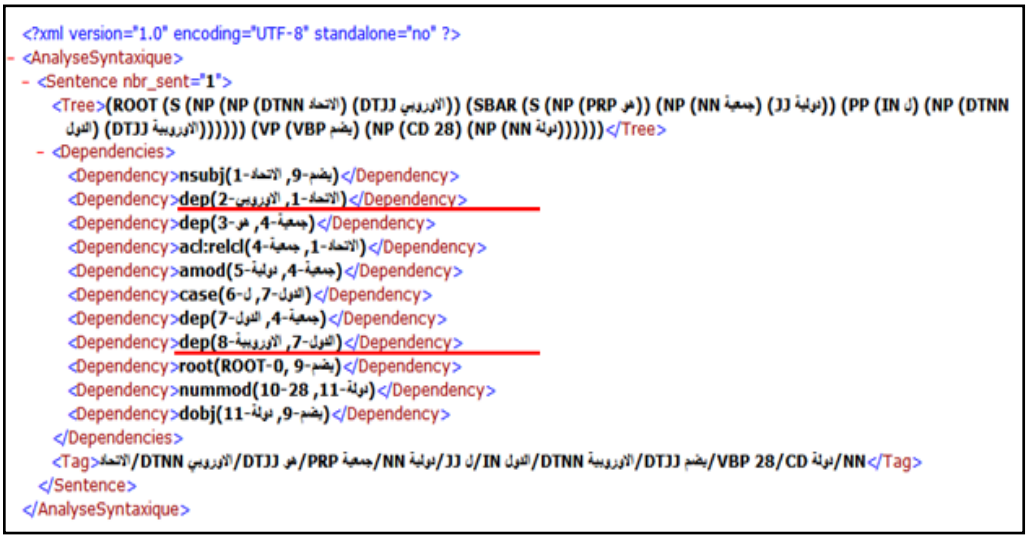


Figure 5.41: Analyse de dépendance fournie par Stanford pour la phrase Phr1

Pour ces raisons, nous proposons d'ajouter une nouvelle règle qui indique si le Gouvernor Tag (GTag) est un nom, le Dependent Tag (DTag) est un adjectif et le type de dépendance retenue par l'analyseur Stanford est " dep ". Alors, l'extraction de relation de la dépendance est construite suivant le modèle:

[Conc(D)]<-(attributeOf)-[Conc(G)]

En appliquant cette règle, nous obtenons la liste de nouvelles relations entre les concepts illustrées dans la figure 5.42.

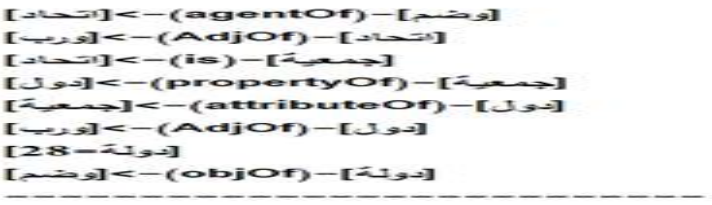


Figure 5.42: Extraction de relations entre concepts en appliquant la règle 1 ajoutée

- ***Règle 2 :***

Soit la phrase Phr2 suivante :" **ذهب الولد الى فرنسا**", en appliquant les règles de [Abouenour, 2014], nous constatons qu'il y a des informations négligées telles que les arguments obliques et les adjoints. Ceci résulte un manque d'informations comme affiche la figure 5.43.

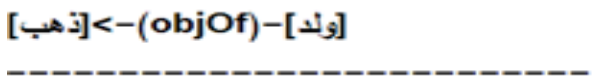


Figure 5.43: Relations entre concepts retenues en appliquant les règles de [Abouenour, 2014]

En consultant le résultat de l'analyse syntaxique fournie par Stanford (figure 5.44), nous trouvons que la relation qui relie " **ذهب** " et "**فرنسا** " est « nmod ». En effet, la relation « nmod » est un nom fonctionnant comme un adjoint ou un argument non-core (oblique)[21]. Ainsi, un argument oblique implique presque toujours un lien existentiel. Alors qu'un adjoint est une partie facultative ou structurellement dispensable d'une phrase, d'une clause ou d'une phrase qui, si elle est enlevée ou rejetée, n'affectera pas le reste de la phrase.

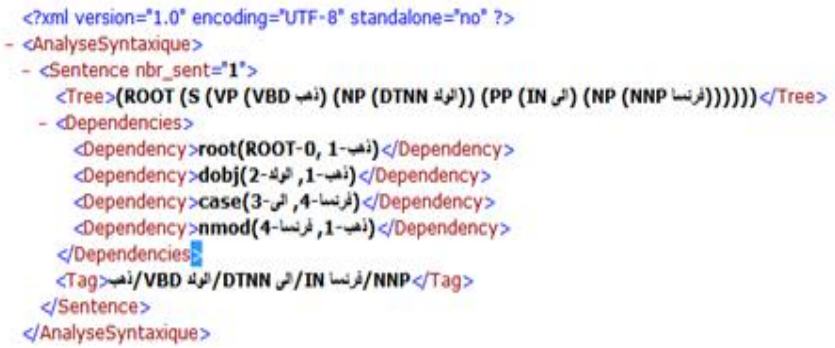


Figure 5.44: Analyse de dépendance fournie par Stanford pour la phrase Phr2

[21] http://gawron.sdsu.edu/semantics/course_core/background/html/logic_lecture/node27.html

Dans ce cadre, nous proposons d'ajouter une nouvelle règle qui indique si le Gouvernor Tag (GTag) est un verbe, le Dependent Tag (DTag) est un nom et le type de dépendance retenue par l'analyseur Stanford est "nmod". Alors, l'extraction de relation de la dépendance est construite suivant le modèle:

[Conc(D)]<-(ArgOf)-[Conc(G)]

En appliquant cette règle, nous obtenons la liste de nouvelles relations entre les concepts illustrées dans la figure 5.45.

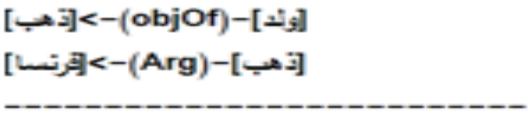


Figure 5.45: Extraction de relations entre concepts en appliquant la règle 2 ajoutée

- ***Autre règles issues lors de l'analyse des questions :***

Pour l'extraction de relations qui relient les concepts nous proposons également de nouvelles règles suivantes en se basant sur le résultat de l'analyse syntaxique de Stanford. Alors, l'extraction de relations de la dépendance est construite pour chaque dépendance suivant ces modèles:

- Si le Dependent Tag (DTag) est égal à "**من**" et le type de dépendance n'est pas égal à "**root**" alors l'extraction de relation de la dépendance est construite suivant le modèle:

[Person : PX] <-(agentOf)-[Conc(G)]

- Si le Dependent Tag (DTag) est égal à "**أين** " et le type de dépendance est égal à "**advmod**" alors l'extraction de relation de la dépendance est construite suivant le modèle:

[Location : LX] <-(LocOF)-[Conc(G)]

- Si le Gouvernor Tag (GTag) est égal à "**ماهي** " ou"**ماهو**" et le type de dépendance n'est pas égal à "**root**" alors l'extraction de relation de la dépendance est construite suivant le modèle:

[Organisation : OX] <-(is)-[Conc(D)]

- Si le Dependent Tag (DTag) est égal à "**متى** " et le type de dépendance est égal à "**advmod**" alors l'extraction de relation de la dépendance est construite suivant le modèle:

[Date : TX] <-(TMP)-[Conc(D)]

- Si le Dependent Tag (DTag) est égal à "**كم** " et le type de dépendance est égal à "**advmod**" alors l'extraction de relation de la dépendance est construite suivant le modèle:

[Numerical_expression : NX] <-(Value)-[Conc(D)]

Autre erreur liée aussi aux règles de [Abouenour, 2014] illustrée surtout en appliquant les deux règles 4 et 5. Avec ces deux règles, le pattern est représenté comme suit : CG-dep = [cg : [SupConc(D) : D]<-objOf-[Conc(G)]]. Nous mentionnons précédemment que SubConc(D) est remplacé par le type d'entité nommée du dépendant. Citons par exemple un cas d'erreur rencontré avec la question suivante « **أين تقع دولة تونس؟** ». Selon les résultats d'ArNER et de Stanford, aucune dépendance contenant le mot « **تونس** » ne satisfait les conditions de la règle 4 ni de la règle 5 (qui sont les seuls contenant « SubConc» dans leurs patterns). Par conséquent, le type d'entité nommée n'est pas pris en considération et n'est pas apparu dans le graphe conceptuel de cette question.

Tout au long du développement de NArQAS, nous avons rencontré des erreurs provenant de l'outil de reconnaissance des entités nommées ArNER. Notons que nous traitons cinq types des questions factuelles mais ArNER ne reconnait que trois types d'entités nommées, à savoir, personne, organisation et location. Dans ce cadre, avec les questions qui commencent par « **متى**» ou « **كم** » lorsque nous cherchons respectivement une entité nommée de type « Date » ou « Expression Numérique », il devient difficile de détecter la réponse précise.

De plus, nous rencontrons des cas où ARNER ne réussit pas à reconnaitre l'entité nommée correspondante. A titre d'exemple, soit une paire de passage-hypothèse représentés respectivement en forme logiques FOLT et FOLH.

Question : أين تقع تونس؟

FOLH : ∃X : Location(LX) ∧ قعي(X) ∧ Loc(X,LX) ∧ Location(تونس) ∧ objOf(X,تونس)

Passage réponse : منظمة عين تونس هي منظمة لرصد مشاغل المواطن والمجتمع وتوجية السلطات التنفيذية و القضائية و التشريعية

FOLT : ∃X ∃Y ∃Z ∃W ∃T ∃E ∃F ∃G ∃H ∃I ∃J : منظمة(X) ∧ عين(Y) ∧ is(X,Y) ∧ تونس(Z) ∧ is(Y,Z) ∧ منظمة(W) ∧ is(W,X) ∧ رصد(T) ∧ is(T,W) ∧ شغل(E) ∧ is(E,T) ∧ مواطن(F) ∧ attributeOf(F,E) ∧ مجتمع(G) ∧ attributeOf(G,F) ∧ توج(H) ∧ is(H,F) ∧ سلطات(I) ∧ attributeOf(I,H) ∧ نفذ(J) ∧ propertyOf(I,J) .

Dans cet exemple, l'expression «**تونس**» est une entité nommée reconnue comme « location » dans l'hypothèse. Cette entité nommée doit être reconnue comme « organisation » dans le passage. Si c'est le cas, l'entité nommée du texte ne correspond pas à celle de l'hypothèse ce qui implique que FOLT n'implique pas le FOLH (figure 5.46). Malheureusement, cette entité n'est pas détectée avec ArNER dans le cas de passage. Ce type d'erreur peut influencer sur la comparaison entre les entités nommées. Spécifiquement, dans l'étape de la RTE, nous pouvons considérer que les entités nommées se concordent or réellement elles se diffèrent.

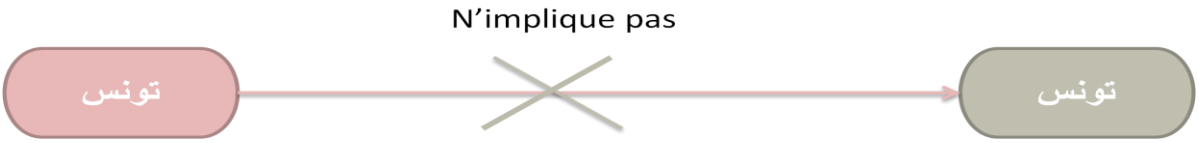


Figure 5.46: Cas de non correspondance entre entités

4.2 Performance de NArQAS en comparaison avec Qwant et Ask.com

Notre approche peut être comparée à plusieurs autres approches. Néanmoins, nous ne connaissons aucun travail antérieur en arabe qui est déjà concentré sur une représentation sémantique et logique, en même temps, et une implication textuelle pour un système de question-réponse, qui répond aux questions factuelles. Par conséquent, nous n'avons donc pas pu comparer nos résultats avec les résultats des systèmes de question-réponse proposés en arabe. Cependant, les moteurs de recherche peuvent ainsi être le premier élément évident de comparaison. En effet, la raison de choisir ces moteurs pour la comparaison est d'avoir une réponse à la question suivante « pourquoi élaborer des systèmes de question-réponse si des moteurs de recherche comme Qwant, Google, Ask.com, et autres, renvoient des réponses correctes ? ». Absolument, notre approche est proche de celles proposées en anglais mais elle est trop éloignée des approches proposées en question-réponse arabe pour permettre une comparaison directe. Ainsi, les systèmes arabes qui sont conçus à trouver les réponses à des questions factuelles ont des architectures très variées.

D'ailleurs, une comparaison de la performance de NArQAS avec un moteur de recherche de référence a illustré la signification du travail. La comparaison s'établit via des mésures d'évaluation, à savoir, la précision, le rappel et la F-mesure. Pour avoir un aperçu clair des performances de NArQAS, nous le comparons avec deux moteurs de recherche Ask.com[22] et Qwant[23]. Clairement, dans le cas de la comparaison, nous dépendons des cinq premières réponses si elles sont correctes, et nous abandonnons les autres réponses trouvées. Nous avons exécuté plusieurs expériences avec les types de questions recueillies dans notre corpus pour obtenir les réponses par Qwant, Ask.com et NArQAS. Dans ce qui suit, nous examinons ces moteurs de recherche et discutons les différences observées pendant la comparaison.

Qwant, est un moteur de recherche lancé en France, a été présenté en 2013 sur le marché après une phase de développement de deux ans. Il a été fondé par Constant Patrick, Jean-Manuel Rozan, Léandri Eric avec la vision de créer un moteur de recherche européen qui démocratise l'information. Qwant propose des services de recherche multiplateformes regroupant et introduisant des informations pertinentes tels que des photos et des vidéos provenant de sites Web, d'actualités et de médias numériques, de sites commerciaux, de réseaux sociaux, etc. Qwant est disponible en 16 langues et accessible par 25 pays différents. La mission de Qwant est de renforcer la confidentialité, la transparence, la confiance, etc., sur Internet ; il est une alternative à Google, Bing, Yahoo, DuckDuckGo, Blekko ou Ghostery.

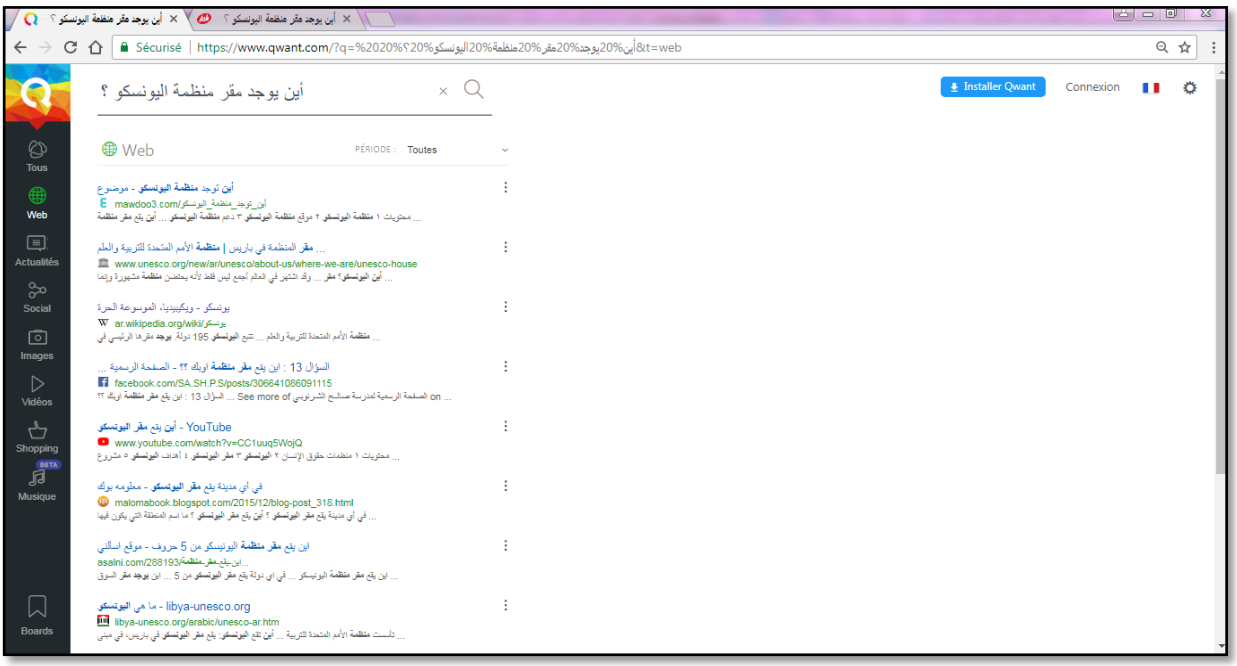


Figure 5.47: Les réponses de Qwant pour la question Q-teste

[22] https://fr.ask.com/?o=0
[23] https://www.qwant.com/

Ask.com est un moteur de recherche américain créé en 1996. Il a été introduit au départ sous le nom d'Ask Jeeves qui avait comme objectif de répondre de manière facile et accessible à des recherches effectuées dans des langues différentes (anglais, français, allemand, espagnol, italien, arabe, etc). Ce moteur utilise des partenaires et possède sa base de connaissances spécifique afin de répondre aux questions. D'ailleurs, il accepte les questions familièrement exprimées et retourne des liens hypertextes vers des pages Web contenant des mots-clés similaires à ceux des questions. L'instantané des réponses du moteur Ask.com à la question Q-Teste, est présenté dans la figure 5.48.

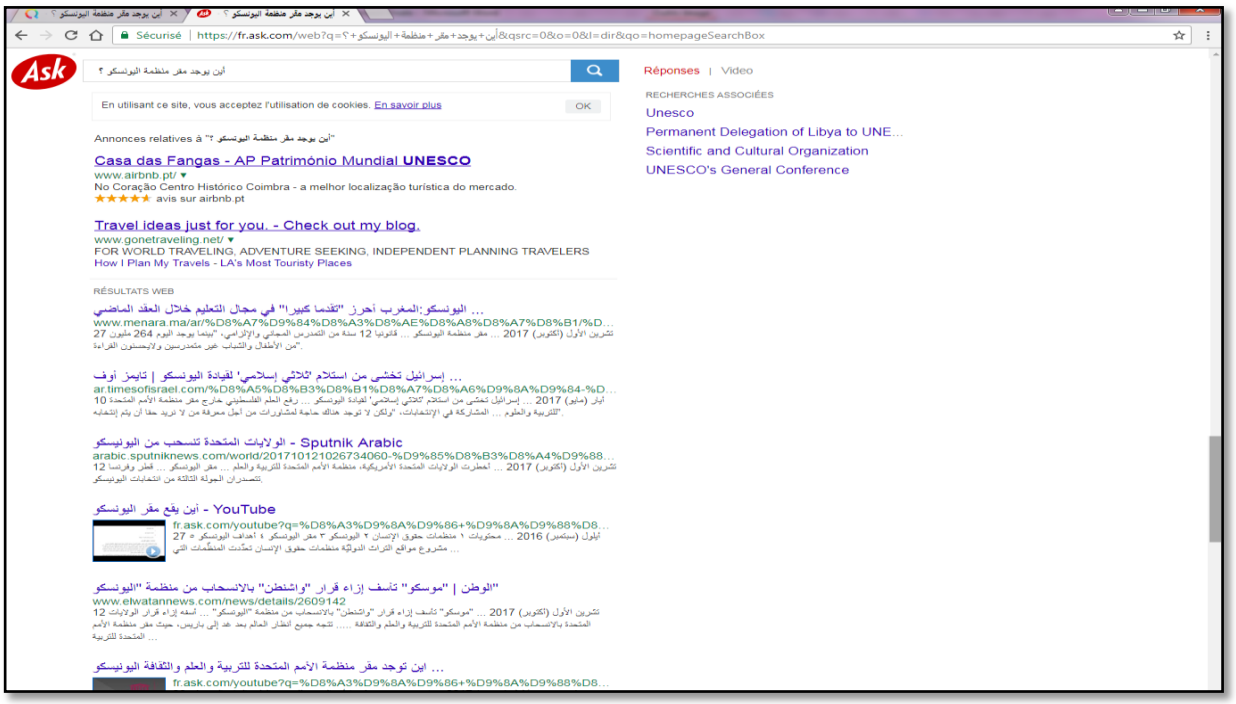


Figure 5.48: Les réponses d'Ask.com pour la question Q-teste

Nous avons réalisé la comparaison sur un ensemble de 20 questions de différents types comme les questions qui seront illustrées dans le tableau 5.28. Nous considérons que l'utilisateur veut une réponse exacte à sa question et nous considérons que la réponse est correcte si cette réponse peut être trouvée dans le passage retenu par NArQAS ou dans le texte lié aux liens hypertextes récupérés par Ask.com ou Qwant. En fonction de l'ensemble des questions avec leurs passages réponses, nous calculons la précision, le rappel ainsi que la F-mesure pour chaque système. Il est important de noter que nous n'avons pas accéder à la base de connaissances de Qwant et Ask .com. D'ailleurs, pour calculer la précision, le rappel et la F-mesure, nous n'avons considéré que les dis premiers liens hypertextes récupérés.

Tableau 5.28: Exemples de questions utilisées pour la comparaison

Exemple de questions	Type de la Question	Type de réponse attendu
من صمم برج ايفل ؟	Who-من	Personne
أين يوجد مقر منظمة اليونسكو ؟	Where-أين	Location
متى استقلت تونس ؟	When-متى	Date
ماهي عاصمة ماليزيا ؟	What ماهو/ماهي	Organisation
كم يبلغ طول نهر الأمازون ؟	How-كم	Expression numérique

Afin de comparer la qualité de ces systèmes, nous avons mesuré le pourcentage de réponses correctes trouvées dans les cinq premiers passages pour NArQAS ou dans les cinq premiers liens hypertextes pour Qwant et Ask. D'ailleurs la précision correspond au pourcentage de questions où la réponse retournée en cinq premières positions est la bonne réponse. Le rappel correspond au pourcentage de questions dont la liste des réponses retournées contient la bonne réponse. La F-mesure est la moyenne harmonique entre le rappel et la précision. Pour cette évaluation nous avons fixé le nombre maximum de réponses retournées à dix passages ou à dix liens hypertextes. Nous présentons ces résultats dans le tableau 5.29.

Tableau 5.29: Performance de NArQAS en comparaison avaec Ask.com et Qwant

Le système	Précision	Rappel	F-mesure
Qwant	0,62	0,59	0,60
Ask	0,67	0,64	0,65
NArQAS	**0,75**	**0,68**	**0,71**

La comparaison entre NArQAS, Qwant et Ask.com en termes de précision, rappel et F-mesure est représentée dans la figure 5.49. Cette figure présente la comparaison de ces systèmes en se basant sur les résultats affichés dans le tableau 5.29. Les résultats démontrent que NArQAS est uniformément plus efficace que les moteurs de recherche Ask.com et Qwant en termes de précision, rappel et F-mesure. Ces résultats prouvent également que notre système peut mieux déterminer la réponse correcte et précise à une question en arabe.

Toutes les expériences menées dans nos travaux illustrent que NArQAS a toujours donné la bonne réponse comme les premières réponses. Ceci peut être attribué à l'utilisation de l'implication textuelle dans le module de la représentation logique et qui est prise en compte dans le module d'extraction de cette réponse. En revanche, Qwant et Ask.com ont généralement fourni la bonne réponse dans le troisième ou le quatrième lien hypertexte.

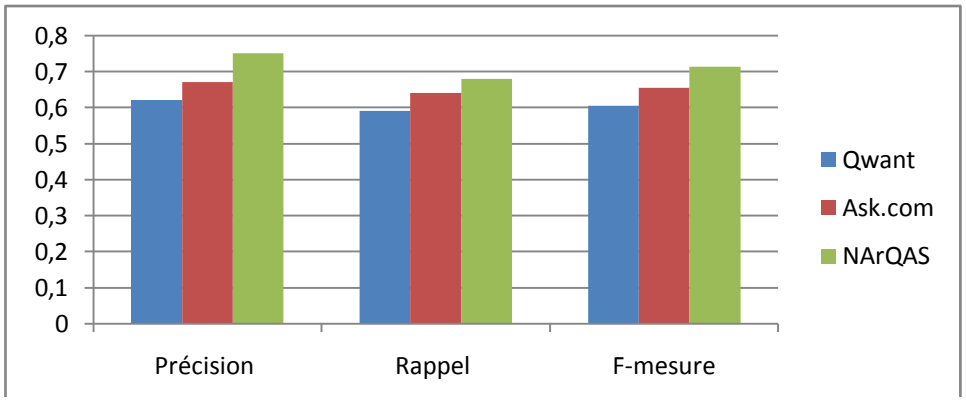


Figure 5.49: Performance de NArQAS en comparaison avec Ask.com et Qwant

Les résultats obtenus démontrent que notre système présente une très bonne performance en termes de précision, rappel et F-mesure. En fait, nous nous attendions à cette bonne performance et nous croyons que cela est principalement dû à la fois à des représentations sémantiques puis logiques des questions et des passages issues du web et à l'intégration l'implication textuelle. Par conséquent, les résultats obtenus montrent que l'utilisation de ces trois technologies (les graphes conceptuels, le raisonnement logique et l'implication textuelle) peut aider de façon significative à aborder le module d'extraction de la réponse précise dans un système de question réponse arabe.

Conclusion

Dans ce chapitre, nous avons présenté la validation de notre approche qui a été illustré à travers un système de question-réponse pour l'arabe «NArQAS». Plus particulièrement, nous avons décrit son architecture, ainsi les principes des différents modules qui constituent cette architecture et le fonctionnement de notre système. NArQAS est doté d'une architecture modulaire et repose sur la combinaison des techniques de recherche d'informations, d'extraction d'informations, de traitement automatique de la langue et de raisonnement automatique. Ainsi, il permettra aux utilisateurs de trouver une réponse précise à une question factuelle. En outre, ce chapitre traite des implémentations en détails pour mettre en œuvre chaque module de cette architecture. Nous avons également présenté, dans ce chapitre, les résultats obtenus par NArQAS et quelques cas d'erreurs rencontrés ainsi que les traitements effectués pour résoudre ces problèmes. A la fin de ce chapitre, nous avons comparé les résultats de notre système par d'autres approches notamment les moteurs de recherche Ask.com et Qwant.

Cette thèse s'achève par une conclusion générale et quelques pistes d'amélioration.

Conclusion générale et perspectives…

CONCLUSION GENERALE

L'objectif principal de cette thèse était de proposer une nouvelle approche pour la question-réponse arabe. Cette approche a rassemblé plusieurs techniques telles que les techniques de TALN, de raisonnement logique, de RTE, etc. Elle est basée essentiellement sur la compréhension automatique de textes arabes (question ou passages de textes) afin de les transformer en des représentations sémantiques et logiques. Elle est conçue également pour déterminer l'implication textuelle entre des paires de représentations logiques des passages (texte) et de la question (hypothèse) afin de trouver le passage de texte qui répond à la question et de sélectionner la réponse précise.

Nous avons tout d'abord présenté une revue de la littérature de différents systèmes de question-réponse proposés dans différentes langues les plus connues telles que l'anglais, le français, etc. Nous avons illustré les principaux travaux existants en termes d'approches et de systèmes dans l'arabe. Nous avons présenté également une analyse performante de ces systèmes. Nous avons décrit quelques recherches qui soulignent la représentation des textes via des graphes conceptuels, de représentations logiques en langues latines et quelques travaux en relation en arabe. Nous avons illustré de nombreux travaux qui ont été menés pour déterminer l'implication textuelle pour la question-réponse en plusieurs langues, y compris l'arabe.

Ensuite, nous avons construit notre corpus de questions-textes AQA-WebCorp en interrogeant le moteur de recherche Google. Pour le faire, nous avons proposé une méthode implémentée en Java. Ainsi, nous avons collecté 250 questions qui peuvent être posées dans différents domaines en occurrence le sport, l'histoire et l'islam, les découvertes et la culture, les nouvelles du monde, la santé et la médecine. La collecte de ces questions a été réalisée à partir de plusieurs sources telles que les forums de discussion, les questions fréquemment posées et quelques questions traduites à partir les deux campagnes d'évaluation TREC et CLEF. Pour chaque question, nous avons récupéré un ensemble de passages de textes. La taille de notre corpus est dans l'ordre de 250 paires de questions-textes dont 50 questions ont été traduites à partir du TREC et du CLEF.

Enfin, nous avons validé notre approche à travers un système de question-réponse implémenté en Java et nommé NarQAS. Ce système fournit aux utilisateurs la possibilité de trouver des réponses précises à leurs questions. Il permet de chercher des réponses à partir du Web à des questions factuelles. En effet, NArQAS est un système complet allant de l'analyse de la question à la génération de la réponse précise. Il est basé sur une analyse sémantique et logique de la question et de ces passages. Ainsi, il extrait des caractéristiques à partir des formes logiques de la question et des passages pour déterminer l'implication textuelle et sélectionner la réponse précise. Ce système est doté d'une architecture modulaire et repose sur la combinaison des techniques de recherche d'informations, d'extraction d'informations, de raisonnement automatique, de traitement automatique de la langue et de reconnaissance d'implications textuelles.

Nous avons réalisé trois expérimentations sur notre corpus. A la première expérience, nous avons évalué NArQAS avec 50 questions, à la deuxième expérience, nous avons utilisé 115 questions pour l'évaluation. Enfin, nous avons accompli les 250 questions. Les résultats parvenus par NArQAS sont encourageants. Pour avoir un aperçu clair des performances de NArQAS, nous l'avons comparé avec les deux moteurs de recherche Qwant et Ask.com. D'après les expériences menées dans cette thèse, nous avons illustré que NArQAS a toujours donné la bonne réponse comme les premières réponses qui peuvent être attribuées à l'utilisation de l'implication textuelle dans le module de la représentation logique et qui sont prises en compte dans le module d'extraction de cette réponse. En revanche, Qwant et Ask.com ont généralement fourni la bonne réponse dans le troisième ou le quatrième lien hypertexte.

L'objectif de cette thèse était non seulement d'obtenir des réponses exactes et précises à des questions en arabe mais de réaliser des analyses sémantiques et logiques pour la question et pour les passages de texte. Pour une représentation logique nous nous sommes référés à une représentation sémantique. L'idée était de transformer ces textes sous forme des graphes conceptuels, un formalisme puissant, fondé sur la logique, qui sert à modéliser les informations textuelles par des concepts et des relations. À partir de ce formalisme, nous avons proposé un algorithme de conversion permettant de modéliser et de déterminer une représentation logique pour chaque graphe. Enfin, nous avons reposé sur l'extraction et la combinaison des caractéristiques pour déterminer l'implication textuelle entre des paires de représentations logiques de la question et du passage.

L'originalité de cette thèse réside dans la proposition d'une nouvelle approche pour la question-réponse en arabe qui a fournit d'une part une analyse sémantique et logique de la question et des textes à partir desquels la réponse a été identifiée. D'autre part, elle a inclus l'implication textuelle à ce domaine d'étude. En conséquence, la modélisation et la mise en œuvre de cette approche est la partie la plus originale de notre contribution, tant d'un point de vue sémantique, logique ou d'un point de vue implication textuelle. D'autre part, l'originalité se manifeste par le développement du système NArQAS. Celui-ci diffère des systèmes existants dans la littérature. La plupart de ces systèmes sont fondés sur des approches morphosyntaxiques. Soulignons que peu de systèmes sont basés sur des approches sémantiques avec très peu sont focalisés sur des approches logiques alors notre système trouve une solution pour prendre en considération les deux à la fois en intégrant la technique de la reconnaissance d'implications textuelles à ce domaine de recherche.

La réalisation de cette thèse ouvre deux types de perspectives, à court terme et à long terme.

Trois perspectives sont envisageables à court terme. La première est l'extension de NArQAS. Le système proposé comporte une étape de détermination d'implication textuelle. Cette étape est assurée par extraction de trois caractéristiques (ex. chevauchement des prédicats-arguments, correspondance des entités nommées et similarité sémantique). Notre première préoccupation dans le futur concerne l'incorporation de nouvelles caractéristiques (ex. des techniques de théorème de démonstration). La deuxième est la finalisation du corpus AQA-WebCorp. Nous pensons étendre notre corpus de questions-textes afin qu'il supporte un échantillon vaste pour une base d'expérimentation future. Nous envisageons également le développement d'une version Web de notre corpus AQA-WebCorp pour des utilisations futures. La troisième est la modification de la méthode de construction de graphes conceptuels. Dans la méthode proposée pour la construction des graphes conceptuels nous avons utilisé des règles proposées par Abouenour afin d'extraire les relations entre concepts. Pour contribuer positivement dans l'amélioration de notre approche, nous proposons de modifier cette méthode.

Les perspectives à long terme concernent d'abord la prise en compte de nouveaux types de questions. Jusqu'à présent, le système ne travaille qu'avec des questions factuelles. Afin d'être le plus exhaustif possible, il devrait s'appliquer à d'autres types de questions. Il

apparaît en conséquence qu'il est utile d'étudier de nouveaux types de questions afin de mieux valoriser la performance de notre système tels que les questions complexes. Finalement, nous envisageons la prise en compte des règles d'inférence pour la reconnaissance d'implication textuelle. Dans nos travaux, l'extraction de caractéristiques est efficace pour déterminer l'implication textuelle. Nous nous intéressons à quelques sujets de recherche qui nous semblent importants au regard de notre travail. Un de ces sujets rejoint en quelque sorte l'application des règles d'inférence pour l'implication textuelle. En effet, l'inférence textuelle joue un rôle important dans de nombreuses tâches de traitement du langage naturel (TALN).

PUBLICATIONS DE L'AUTEUR

Des publications inspirées par nos travaux de thèse, et c'est ce sur quoi nous allons conclure, sont en cours. Ce sont des articles qui font parties des résultats obtenus au cours de nos travaux de recherche. Par conséquent, au moment de l'écriture, les apports de cette thèse ont donné lieu aux plusieurs publications dans des conférences et/ou des journaux spécialisés:

- ***Publications dans des revues internationales***

1. Bakari, W., Bellot, P., & Neji, M. (2017). ***A logical representation of Arabic questions toward automatic passage extraction from the Web***. International Journal of Speech Technology, 20(2), 339-353. (**Indéxée: Scopus, Dblp, *Springer*, IF=1.08, Classe: Q2).**

2. W. Bakari, P. Bellot, M. Neji, ***A Preliminary Study for Building an Arabic Corpus of Pair Questions-texts from the Web: AQA-WebCorp***, iJES 4(2): 38-45 (2016). (**Indéxée: Dblp).**

3. Bakari, W., Bellot, P., Trigui, O., & Neji, M. (2015). «***Towards Logical Inference for Arabic Question-Answering***». Research in Computing Science, 90, 87-99. **(Indéxée: Dblp, Springer , LatIndex, Periodica).**

- ***Publications dans des notes de lecture***

4. W. Bakari, P. Bellot, M. Neji «***Literature review of Arabic Question-Answering: Modeling, Generation, Experimentation and performance analysis***», 2015 Flexible Query Answering Systems, 26-28 October 2015, Cracow, Poland. **(Indéxée: Scopus, Dblp, Springer, Classe: C).**

- ***Publications dans des conférences internationales***

5. W. Bakari, P. Bellot, M. Neji (2017, December). ***Generating semantic and logic meaning representations when analyzing the Arabic natural questions.*** In International Conference on Intelligent Systems Design and Applications. **(Indéxée: ISI Proceedings, Dblp, Scopus, Springer, Classe: C).**

6. M. Ben-Sghaier, W. Bakari, M. Neji (2017, December). ***An Arabic question-answering system combining a semantic and logical representation of texts***. In International Conference on Intelligent Systems Design and Applications. **(Indéxée: ISI Proceedings, Dblp, Scopus, Springer, Classe: C).**

7. W. Bakari, P. Bellot, M. Neji « ***Using the web as an efficient source of building an Arabic corpus: presentation and evaluation*** », 2016 27th IBIMA International Conference, 4 - 5 May 2016 Milan, Italy. **(Indéxée: ISI Thomson reuters, Scopus, Elsevier, Classe: B).**

8. W. Bakari, P. Bellot, M. Neji « ***AQA-WebCorp: Web-based Factual Questions for Arabic*** », 2016 20th 20th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, 5, 6 & 7 September 2016 York, UK. (**Indéxée: *ISI proceedings, Dblp, Elsevier, Scopus, Science direct,*, Classe: B).**

9. W. Bakari, P. Bellot, M. Neji « ***Researches and Reviews in Arabic Question Answering: principal approaches and systems with classification***», International Arab Conference on Information Technology (ACIT'2016), Morocco, Beni-Mellal, 6-8, December 2016.

10. W. Bakari, O. Trigui, M. Neji« ***Logic-based approach for improving Arabic question answering*** », 2014 IEEE International Conference On Computational Intelligence And Computing Research, December 18-20, 2014, Tamilnadu, India. **(Indéxée: Scopus, IEEE).**

BIBLIOGRAPHIE

[Abdelnasser et al., 2014] Abdelnasser, H., Ragab, M., Mohamed, R., Mohamed, A., Farouk, B., El-Makky, N., & Torki, M. (2014). Al-Bayan: an arabic question answering system for the holy quran. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)(pp. 57-64).

[Abouenour et al., 2014] Abouenour, L., Nasri, M., Bouzoubaa, K., Kabbaj, A., & Rosso, P. (2014). Construction of an ontology for intelligent Arabic QA systems leveraging the Conceptual Graphs representation. Journal of Intelligent & Fuzzy Systems, 27(6), 2869-2881.

[Abouenour et al.,2012] Abouenour, L., Bouzoubaa, K., & Rosso, P. (2012). IDRAAQ: New Arabic Question Answering System Based on Query Expansion and Passage Retrieval. In CLEF (Online Working Notes/Labs/Workshop).

[Abouenour, 2014] Abouenour, L. (2014). Three-levels Approach for Arabic Question Answering Systems. Diss. Ecole Mohammadia d'Ingénieurs.

[Abufardeh & Magel, 2008] Abufardeh, S. and K. Magel, 2008. Software localization: The Challenging Aspects of Arabic to the Localization Process (Arabization). IASTED Proceeding of the Software Engineering SE 2008, Innsbruck, Austria, pp: 275-279.

[Abuleil & Evens, 2002] Abuleil S., Evens M. (2002) Extracting an Arabic Lexicon from Arabic Newspaper Text. Computers and the Humanities, 36(3), pp. 191-221.

[Agirre et al., 2010] Eneko Agirre, Arantxa Otegi, Hugo Zaragoza, Using semantic relatedness and word sense disambiguation for (CL) IR, in: Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peas, Giovanna Roda (Eds.), Multilingual Information Access Evaluation I. Text Retrieval Experiments, Lecture Notes in Computer Science, vol. 6241, Springer, Berlin/Heidelberg, 2010, pp. 166–173.

[Akour et al., 2011] Akour M., Abufardeh S., Magel K. and Al-Radaideh Q.," QArabPro: A Rule Based Question Answering System for Reading Comprehension Tests in Arabic", American Journal of Applied Sciences, vol. 8 (6), pp. 652-661, 2011.

[Alabbas & Ramsay, 2013] Alabbas, M., & Ramsay, A. (2013). Natural language inference for Arabic using extended tree edit distance with subtrees. Journal of Artificial Intelligence Research, 48, 1-22.

[Alabbas, 2011] M. Alabbas, "ArbTE: Arabic Textual Entailment," in Proceedings of the 2nd Student Research Workshop associated with RANLP, pp. 48-53, 2011.

[Alagha & Abu-Taha, 2015] AlAgha I, Abu-Taha A (2015) AR2SPARQL: An Arabic Natural Language Interface for the Semantic Web, International Journal of Computer Applications, 125(6), pp 19-27.

[Al Chalabi, 2015] Al Chalabi, H. M. (2015). Question Processing for Arabic Question Answering System (Doctoral dissertation, The British University in Dubai (BUiD)).

[Al-daimi & Abdel-amir, 1994] Al-daimi, K., M. Abdel-amir, 1994. The syntactic analysis of arabic by machine. Comput. Humanities, 28: 29-37.

[Al-Khalifa & Al-Wabil, 2007] Al-Khalifa H, Al-Wabil A (2007) The Arabic language and the semantic web: Challenges And opportunities, In: The first international symposium on computers and the Arabic language, November 2007, Riyadh, Saudi Arabia.

[AL-Khawaldeh, 2015] AL-Khawaldeh, F. T. (2015). A Study of the Effect of Resolving Negation and Sentiment Analysis in Recognizing Text Entailment for Arabic. World of Computer Science and Information Technology Journal (WCSIT), 5(7), 124-128.

[Allam & Haggag, 2012] Allam, A. M. N., & Haggag, M. H. (2012). The question answering systems: A survey. International Journal of Research and Reviews in Information Sciences (IJRRIS), 2(3).

[Almarwani & Diab, 2017] Almarwani, N., & Diab, M. (2017). Arabic Textual Entailment with Word Embeddings. WANLP 2017 (co-located with EACL 2017), 185–190

[Athenikos & Han, 2010] Athenikos SJ., Han H. (2010).“Biomedical question answering: A survey”, Computer Methods and Programs in Biomedecine 99 (1):24, PMID, 19913938.

[Ayache, et al. 2006] Ayache, C., Grau, B., & Vilnat, A. (2006). EQueR: the French Evaluation campaign of Question-Answering Systems. collections, 2(2), 2-3.

[Babych & Hartley, 2003] Babych, B., & Hartley, A. (2003, April). Improving machine translation quality with automatic named entity recognition. In Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT (pp. 1-8). Association for Computational Linguistics.

[Bakari et al., 2014] Bakari, W., Trigui, O., and Neji, M. (2014, December). Logic-based approach for improving Arabic question answering. In Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on (pp. 1-6). IEEE.

[Bakari et al., 2015] Bakari, W., Bellot, P., & Neji, M. (2015). Literature Review of Arabic Question-Answering: Modeling, Generation, Experimentation and Performance Analysis. In Flexible Query Answering Systems 2015 (pp. 321-334). Springer International Publishing.

[Bakari et al., 2016] Bakari, W., Bellot, P., & Neji, M. (2016). AQA-WebCorp: Web-based factual questions for Arabic. Procedia Computer Science, 96, 275-284. ISO 690

[**Bakari et al., 2017]** W. Bakari, P. Bellot, M. Neji (2017, December). Generating semantic and logic meaning representations when analyzing the Arabic natural questions. In International Conference on Intelligent Systems Design and Applications. ISI Proceedings, dblp, Scopus, Springer, Cham.

[Banerjee et al., 2013] Banerjee, S., Bhaskar, P., Pakray, P., Bandyopadhyay, S., & Gelbukh, A. F. (2013, September). Multiple Choice Question (MCQ) Answering System for Entrance Examination. In CLEF (Working Notes).

[Baral et al., 2005] Baral, C., Gelfond, G., Gelfond, M., Scherl, R.B., (2005), Textual inference by combining Multiple Logic Programming paradigms, in: Proceedings of the AAAI 2005 Workshop Inference for Textual Question Answering, AAAI Press, Pittsburgh, PA, USA, 2005, pp. 1–5.

[Barbier, 2009] Barbier, V. (2009). Utilisation de connaissances sémantiques Pour l’analyse de justifications de réponses à des questions(Doctoral dissertation, Université Paris Sud-Paris XI).

[Bauer & Berleant, 2012] Bauer, M. A., & Berleant, D. (2012). Usability survey of biomedical question answering systems. Human genomics, 6(1), 1.

[Bekhti & Al-Harbi, 2013] BEKHTI, S., & AL-HARBI, M. (2013). AQuASys: A question-answering system for Arabic. In WSEAS International Conference. Proceedings. Recent Advances in Computer Engineering Series (No. 12). WSEAS.

[Bélanger, 2006] Bélanger, L. (2006). Architecture question-réponse pour l'automatisation des services d'information. Université de Montréal.

[Belguith et al., 2005] Belguith, L., Baccour, L., & Mourad, G. (2005). Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. In Actes de la 12éme Conférence annuelle sur le Traitement Automatique des Langues Naturelles (pp. 451-456).

[Belguith et al., 2007] Belguith, L. H., Aloulou, C., & Hamadou, A. B. (2007). MASPAR: De la segmentation à l'analyse syntaxique de textes arabes. CÉPADUÈS-Editions, editeur, Revue Information Interaction Intelligence I, 3, 9-36.

[Ben Abacha, 2012] Ben-Abacha, A., "Recherche de réponses précises à des questions médicales : le système de questions-réponses MEANS", PhD thesis. Universite PARIS-SUD 11 LIMSI-CNRS. JUIN 2012.

[Ben-Abacha, 2009] Ben-Abacha, A., Questions-réponses dans le domaine médical: une approche sémantique, MajecSTIC 2009 Avignon, France, du 16 au 18 novembre 2009.

[Benajiba et al.,2007] Benajiba, Y., Rosso, P., & Lyhyaoui, A. (2007, April). Implementation of the ArabiQA question answering system's components. In Proc. Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morroco, April (pp. 3-5).

[Benamara, 2004] Benamara Farah, 2004. Cooperative question answering in restricted domains: the WEBCOOP experiments. In: In Workshop on Question Answering in Restricted Domains, 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, p. 31–38.

[Bernard et al., 2009] Guillaume Bernard, Sophie Rosset, Olivier Galibert, Eric Bilinski, and Gilles Adda. 2009. The limsi participation to the qast 2009 track. In Working Notes for the CLEF 2009 Workshop, Corfu, Greece, October.

[Bernard, 2011] Bernard, G. (2011). Réordonnancement d'hypothèses dans un systèmes de questions réponses (Doctoral dissertation, Ph. D. thesis, Université Paris Sud).

[Besançon et al., 2007] Besançon, R., Embarek, M., & Ferret, O. (2007). Finding Answers in the Œdipe System by Extracting and Applying Linguistic Patterns. Evaluation of Multilingual and Multi-modal Information Retrieval, 395-404.

[Bhaskar et al., 2012] Bhaskar, P., Pakray, P., Banerjee, S., Banerjee, S., Bandyopadhyay, S., & Gelbukh, A. (2012, September). Question Answering System for QA4MRE@CLEF 2012. In CLEF 2012 Workshop on Question Answering For Machine Reading Evaluation (QA4MRE).

[Bilotti & Nyberg, 2008] Bilotti, M. W., & Nyberg, E. (2008, August). Improving text retrieval precision and answer accuracy in question answering systems. In Coling 2008: Proceedings of

the 2nd workshop on Information Retrieval for Question Answering (pp. 1-8). Association for Computational Linguistics.

[Black et al., 2006] Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006, January). Introducing the Arabic wordnet project. In Proceedings of the third international WordNet conference (pp. 295-300).

[Bleik et al., 2010] Bleik, S., Xiong, W., Wang, Y., & Song, M. (2010, December). Biomedical concept extraction using concept graphs and ontology-based mapping. In Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on(pp. 553-556). IEEE.

[Bouhriz et al., 2015] Bouhriz, N., Benabbou, F., & Benlahmer, H. (2015). Text Concepts Extraction based on Arabic WordNet and Formal Concept Analysis International Journal of Computer Applications (0975 – 8887) Volume 111 – No 16, February.

[Brill et al., 2002] Brill, E., Dumais, S., and Banko, M. (2002, July). An analysis of the AskMSR question-answering system. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 257-264). Association for Computational Linguistics.

[Brini et al.,2009] Brini, W., Ellouze, M., & Hadrich Belguith, L. (2009). QASAL: Un système de question-réponse dédié pour les questions factuelles en langue Arabe. 9ème Journées Scientifiques des Jeunes Chercheurs en Génie Electrique et Informatique, Tunisia.

[Burger et al., 2001] Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S.M., (2001). Issues, tasks and program structures to roadmap research in question & answering (Q & A). Rapport technique, NIST.

[Cabrio et al., 2012] Cabrio, E., Cojan, J., Aprosio, A. P., Magnini, B., Lavelli, A., & Gandon, F. (2012, November). QAKiS: an open domain QA system based on relational patterns. In Proceedings of the 2012th International Conference on Posters & Demonstrations Track-Volume 914 (pp. 9-12). CEUR-WS. org.

[Cao et al., 2010] Cao, X., Cong, G., Cui, B., & Jensen, C. S. (2010, April). A generalized framework of exploring category information for question retrieval in community question answer archives. In Proceedings of the 19th international conference on World wide web (pp. 201-210). ACM.

[Cao et al., 2011] Cao Y., Liu F., Simpson P., Antieau L., Bennett A., Cimino JJ., Ely J. and Yu H. (2011) "AskHERMES: an online question answering system for complex clinical questions", J Biomed Inform, 44(2):277–288.

[Chavan & Gore, 2016] Chavan, G., & Gore, S. (2016). Design of the Effective Question Answering System by Performing Question Analysis using the Classifier. International Journal of Computer Applications, 139(14).

[Clark et al., 2005] Clark, C., Hodges, D., Stephan, J., Moldovan, D., (2005), Moving QA towards reading comprehension using context and default reasoning, in: Proceedings of the AAAI 2005 Workshop Inference for Textual Question Answering, AAAI Press, Pittsburgh, PA, USA, 2005, pp. 6–12.

[Clark et al., 2008] Clark, P., Harrison, P. Recognizing Textual Entailment with Logical Inference. In Proceedings of 2008 Text Analysis Conference (TAC'08), Gaithsburg, Maryland, 2008.

[Clark et., 2012] Clark, P., Harrison, P., & Yao, X. (2012). An Entailment-Based Approach to the QA4MRE Challenge. In CLEF (Online Working Notes/Labs/Workshop).

[Dagan & Glickman, 2004] Dagan, I., and Glickman, O. 2004. Probabilistic textual entailment: generic applied modeling of language variability, In PASCAL Workshop on Learning Methods for Text Understanding and Mining, Grenoble, France.

[Dagan et al., 2005] DAGAN I., GLICKMAN O. & MAGNINI B. (2005). The PASCAL Recognising Textual Entailment Challenge. In Machine Learning Challenges Workshop, p. 177–190.

[Dagan et al., 2006] Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL recognising textual entailment challenge. In Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment (pp. 177-190). Springer Berlin Heidelberg.

[Dagan et al., 2009] Dagan, I., Dolan, B., Magnini, B., & Roth, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. Nat. Lang. Engineering, 15(4), i–xvii. Editorial of the special issue on Textual Entailment.

[Dagan et al., 2013] Dagan, I., Roth, D., Sammons, M., & Zanzotto, F. M. (2013). Recognizing textual entailment: Models and applications. Synthesis Lectures on Human Language Technologies, 6(4), 1-220.

[Dao & Simpson, 2005] Dao, T. N., & Simpson, T. (2005). Measuring similarity between sentences. WordNet. Net, Tech. Rep.

[Demner-Fushman & Lin, 2007] Dina Demner-Fushman, Jimmy J. Lin, Answering clinical questions with knowledge-based and statistical techniques, Computational Linguistics 33 (1) (2007) 63–103.

[Dhanjal & Sharma, 2015] Dhanjal, G. S., and Sharma, S. (2015), advancements in question answering systems towards indic languages. International Journal of Research in Computer Science, 5(1), 15.

[Dornescu, 2010] Iustin Dornescu, Semantic QA for encyclopaedic questions: EQUAL in GikiCLEF, in: Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Penas, Giovanna Roda (Eds.), Multilingual Information Access Evaluation I. Text Retrieval Experiments, Lecture Notes in Computer Science, vol. 6241, Springer, Berlin/Heidelberg, 2010, pp. 326–333.

[Dwivedi, 2013] Dwivedi, S.K., 2013. Research and reviews in question answering system. In: International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA).

[Edward, 1976] Edward H. Shortliffe, Computer Based Medical Consultations: MYCIN, American Elsevier, 1976.

[Ehrmann, 2008] Ehrmann M. (2008). Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation. PhD thesis, Université Paris 7.

[El Ayari et al., 2009] El Ayari, S., & Grau, B. (2009, April). A Framework of Evaluation for Question-Answering Systems. In ECIR (pp. 744-748).

[El Ayari, 2007] El Ayari, S. (2007). Évaluation transparente de systèmes de questions-réponses: application au focus. Actes de ReciTAL.

[Elkateb et al. 2006] Elkateb S., Black W., Vossen P., Farwell D., Rodríguez H., Pease A., Alkhalifa M., "Arabic WordNet and the Challenges of Arabic", In proceedings of Arabic NLP/MT Conference, London, U.K, 2006.

[Embarek & Ferret, 2010] Embarek, M., & Ferret, O. (2010, April). Can Esculape cure the complex of Oedipe in the medical domain?. In Adaptivity, Personalization and Fusion of Heterogeneous Information (pp. 20-23). LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

[Embarek, 2008] Embarek, M. (2008). Un système de question-réponse dans le domaine médical: le système Esculape (Doctoral dissertation, Université Paris-Est).

[Ezzeldin & Shaheen, 2012] Ezzeldin A. M. and Shaheen M. (2012). ―A survey of Arabic question answering: challenges, tasks, approaches, tools, and future trends‖, the 13th International Arab Conference on Information Technology ACIT'2012. Dec.10-13. ISSN 1812-0857.

[Ezzeldin et al., 2013] Ezzeldin, A. M., Kholief, M. H., & El-Sonbaty, Y. (2013, September). ALQASIM: Arabic language question answer selection in machines. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 100-103). Springer, Berlin, Heidelberg.

[Fader et al., 2014] Fader, A., Zettlemoyer, L., & Etzioni, O. (2014, August). Open question answering over curated and extracted knowledge bases. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1156-1165). ACM.

[Falco, 2014] Falco, M. H. (2014). Répondre à des questions à réponses multiples sur le Web (Doctoral dissertation, Université Paris Sud-Paris XI).

[Fellbaum, 1998] Fellbaum, C., (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[Ferrández et al., 2009] Ferrández, O., R. Izquierdo, S. Ferr´andez, ´ and J. L. Vicedo. 2009. Addressing ontology-based question answering with collections of user queries. IPM, 45(2):175–188.

[Ferret et al,. 2002a] Ferret, Olivier, Grau, Brigitte, Hurault-Plantet, Martine, Illouz, Gabriel et Jacquemin, Christian, Quand la réponse se trouve dans un grand corpus, dans Revue d'Ingénierie des Systèmes d'Information, tm. 7(1-2), 2002a, pp. 95–123.

[Ferret et al., 2000] Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Jacquemin, C., Masson, N., & Lecuyer, P. (2000, June). QALC--The Question-Answering System of LIMSI-CNRS. In TREC.

[Ferret et al., 2001a] O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, and C. Jacquemin. 2001a. Document selection refinement based on linguistic features for qalc, a question answering system. In Proceedings of RANLP2001.

[Ferret et al., 2001b] O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, L. Monceaux, I. Robba, and A. Vilnat. 2001b. Finding an answer based on the recognition of the question focus. In Proceedings of TREC 10.

[Ferrucci et al., 2010] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., ... & Schlaefer, N. (2010). Building Watson: An overview of the DeepQA project. AI magazine, 31(3), 59-79.

[Ferrucci, 2012] Ferrucci, D. A. (2012). Introduction to "this is watson". IBM Journal of Research and Development, 56(3.4), 1-1.

[Forner et al., 2008] Forner, P., Peñas, A., Agirre, E., Alegria, I., Forăscu, C., Moreau, N., ... & Sutcliffe, R. (2008, September). Overview of the clef 2008 multilingual question answering track. In Workshop of the Cross-Language Evaluation Forum for European Languages(pp. 262-295). Springer, Berlin, Heidelberg.

[Fowler et al., 2005] Abraham Fowler, Bob Hauser, Daniel Hodges, Ian Niles, Adrian Novischi, & Jens Stephan. Applying cogex to recognize textual entailment. In In Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment, pages 69–72, 2005.

[Frank et al., 2007] Anette Frank, Hans-Ulrich Krieger, Feiyu Xu, Hans Uszkoreit, Berthold Crysmann, Brigitte Jorg, Ulrich Schafer, Question answering from structured knowledge sources, Journal of Applied Logic 5 (1) (2007) 20–48.

[Galibert, 2009] Galibert, O. (2009). Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert (Doctoral dissertation, Université Paris Sud-Paris XI).

[Galibert, et al., 2005] Olivier Galibert, Gabriel Illouz and Sophie Rosset. 2005. Ritel: an open-domain, human-computer dialog system. In: Proceedings of InterSpeech 2005, Lisbon, Portugal, pp. 909-912.

[Ganesh & Varma, 2009] Surya Ganesh and Vasudeva Varma. 2009. Exploiting the use of prior probabilities for passage retrieval in question answering. In RANLP-2009, pages 99– 102, Borovets, Bulgaria, September. Association for Computational Linguistics.

[Garcia-fernandez, 2010] Garcia-Fernandez, A. (2010). Génération de réponses en langue naturelle orales et écrites pour les systèmes de question-réponse en domaine ouvert (Doctoral dissertation, Université Paris Sud-Paris XI).

[Gatto, 2011] Gatto, M., (2011), The _body'and the _web': The web as corpus ten years on.ICAME JOURNAL, 2011, vol. 35, p. 35-58.

[Giampiccolo et., 2007] Giampiccolo, D., Forner, P., Herrera, J., Peñas, A., Ayache, C., Forascu, C., ... & Sutcliffe, R. (2007, September). Overview of the CLEF 2007 multilingual question answering track. In Workshop of the Cross-Language Evaluation Forum for European Languages (pp. 200-236). Springer, Berlin, Heidelberg.

[Glickman et al., 2006] Glickman, O., Dagan, I., Keller, M., Bengio, S., & Daelemans, W. (2006, June). Investigating lexical substitution scoring for subtitle generation. In Proceedings of the Tenth Conference on Computational Natural Language Learning (pp. 45-52). Association for Computational Linguistics.

[Global-research, 2001] Global research global internet statistics, http://www.euromktg.com/globstats,2001.

[Glöckner & Pelzer, 2008] Glockner, I. and Pelzer, B. (2008). Exploring robustness ¨ enhancements for logic-based passage filtering. In Knowledge Based Intelligent Information and Engineering Systems (Proc. of KES2008, Part I), LNAI 5117, pages 606–614. Springer.

[Glöckner, 2007] Glöckner, I. (2007). Filtering and fusion of question-answering streams by robust textual inference. In Proceedings of KRAQ (Vol. 7, pp. 43-48).

[Glöckner & Pelzer, 2010] Ingo Gloeckner, Bjoern Pelzer, Extending a logic-based question answering system for administrative texts, in: Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Penas, Giovanna Roda (Eds.), Multilingual Information Access Evaluation I. Text Retrieval Experiments, Lecture Notes in Computer Science, vol. 6241, Springer, Berlin/Heidelberg, 2010, pp. 265–272.

[Gomez-Adorno et al., 2013] Gómez-Adorno, H., Pinto, D., & Ayala, D. V. (2013). Semantic Answer Validation in Question Answering Systems for Reading Comprehension Tests. In AMW.

[Gómez-Adorno et al., 2014] Gómez-Adorno, H., Sidorov, G., Pinto, D., & Gelbukh, A. F. (2014). Graph Based Approach for the Question Answering Task Based on Entrance Exams. In CLEF (Working Notes) (pp. 1395-1403). ISO 690

[Grappy & Grau, 2010] Grappy, A., & Grau, B. (2010, April). Answer type validation in question answering systems. In Adaptivity, Personalization and Fusion of Heterogeneous Information (pp. 9-15). Le Centre De Hautes Etudes Internationales D'informatique Documentaire.

[Grappy & Grau, 2011] Grappy, A., & Grau, B. (2011). Validation du type de la réponse dans un système de questions réponses. Document numérique, 14(2), 125-147.

[Grappy et al., 2011] Grappy A., Grau B., Falco M.-H., Ligozat A.-L., Robba I., Vilnat A., « Selecting answers to questions from Web documents by a robust validation process », WI, 2011.

[Grappy, 2011] Grappy, A. (2011). Validation de réponses dans un système de questions réponses (Doctoral dissertation, Université Paris Sud-Paris XI).

[Grau & Chevallet, 2008] B. Grau et J.-P. Chevallet, 2008. La recherche d'informations précises : apprentissage, traitement automatique de la langue et connaissances pour les systèmes de questionréponse. Paris : Hermès.

[Grau et al., 2005] Grau, B., Illouz, G., Monceaux, L., Paroubek, P., Pons, O., Robba, I., & Vilnat, A. (2005, January). FRASQUES, le système du groupe LIR, LIMSI. In Atelier EQueR, Conférence (TALN'05).

[Grau et al., 2006] Grau, B., Ligozat, A. L., Robba, I., Vilnat, A., Bagur, M., & Séjourné, K. (2006). The bilingual system MUSCLEF at QA@ CLEF 2006. In Evaluation of Multilingual and Multi-modal Information Retrieval (pp. 454-462). Springer Berlin Heidelberg.

[Grau et al., 2006a] Grau B. & Ligozat A. & Robba I. & Vilnat A. & Monceaux L. - FRASQUES : A quesionanswering system in the EQueR evaluation campaign. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), 2006a

[Grau et al., 2012] Grau B., Pho V-M., Ligozat A-L., Ben Abacha A., Zweigenbaum P. and Chowdhury Md.F.M. (2012). ―Adaptation of LIMSI's QALC for QA4MRE‖, CLEF (Online Working Notes/Labs/Workshop).

[Green et al, 1961] Green Jr, Bert F., et al. "Baseball: an automatic question-answerer." Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference. ACM, 1961.

[Guo, 2004] Guo, Y. (2004). Chinese Question Answering with Full-Text Retrieval Re-Visited (Doctoral dissertation, University of Waterloo).

[Gupta & Gupta , 2012] Gupta, P., & Gupta, V. (2012). A survey of text question answering techniques. International Journal of Computer Applications, 53(4).

[Habash & Rambow, 2005] Habash, N. and O. Rambow, 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp:573-580, June 25- 30, Ann Arbor, Michigan.

[Habash et al., 2009] Habash, N., Rambow, O., Roth, R.: MADA+TOKAN (2009). "A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, PoS Tagging, Stemming and Lemmatization". In: Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, pp. 102–109 (2009)

[Hammo et al., 2004] Hammo, B., Abuleil, S., Lytinen, S., & Evens, M. (2004). Experimenting with a question answering system for the Arabic language. Computers and the Humanities, 38(4), 397-415.

[Harabagiu & Hickl, 2006] Harabagiu, S., & Hickl, A. (2006, July). Methods for using textual entailment in open-domain question answering. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 905-912). Association for Computational Linguistics.

[Harabagiu et al, 2003] Harabagiu, S. M., Maiorano, S. J., & Paşca, M. A. (2003). Open-domain textual question answering techniques. Natural Language Engineering, 9(3), 231-267.

[Harabagiu et al., 2000] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus and P. Morarescu (2000). FALCON: Boosting knowledge for question answering. In Proceedings of the Ninth Text REtrieval Conference (TREC-9).

[Harabagiu, et al. 1999] Harabagiu, S. M., Miller, G. A., & Moldovan, D. I. (1999). Wordnet 2-a morphologically and semantically enhanced resource. SIGLEX99: Standardizing Lexical Resources.

[Hasan, 2008] Hasan, I. A. H. (2008). Alimentation automatique d'une base de connaissances à partir de textes en langue naturelle. Application au domaine de l'innovation (Doctoral dissertation, Université Blaise Pascal-Clermont-Ferrand II).

[Hensman & Dunnion, 2004] Hensman, S., Dunnion, J., 2004. Automatically building conceptual graphs using VerbNet and WordNet. In: Proceedings of the 2004 International Symposium on information and Communication Technologies, Las Vegas, NV, June 16–18, 2004. ACM International Conference Proceeding Series, vol. 90. Trinity College, Dublin, pp. 115–120.

[Hirschman & Gaizauskas, 2001] L. Hirschman, R. Gaizauskas, Natural language question answering: The view from here, Natural Language Engineering 7 (4) (2001) 275–300.

[Hirschman et al, 1999] Hirschman, L., Light M., Breck E. & Burger J.(1999). Deep Read: A reading comprehension system, In Proceedings 37th Annual Meeting of the Association for Computational Linguistics, pp.325-332.

[Hixon et al., 2015] Hixon, B., Clark, P., & Hajishirzi, H. (2015, May). Learning Knowledge Graphs for Question Answering through Conversational Dialog. In HLT-NAACL (pp. 851-861).

[Hovy et al., 2000] Hovy, E., Gerber, L., Hermjacob, U., Junk, M., Lin, C., 2000. Question answering in webclopedia. In: Ninth Text REtrieval Conference, Volume 500–249 of NIST Special Publication, Gaithersburg, MD, National Institute of Standards and Technology, pp. 655–664

[Isaac et al., 2001] Issac, F., Hamon, T., Bouchard, L., Emirkanian, L., and Fouqueré, C., (2001), extraction informatique de données sur le web : une expérience, in Multimédia, Internet et francophonie : à la recherche d'un dialogue, Vancouver, Canada, mars 2001.

[Ismail et al., 2013] Ismail, S., Moawd, I., & Aref, M. (2013). Arabic text representation using rich semantic graph: A case study. In Proceedings of the 4th European conference of computer science (ECCS'13) (pp. 148-153).

[Jacquemin, 1999] Jacquemin C., (1999), Syntagmatic and paradigmatic representations of term variation, Actes de ACL'99, 341-348.

[Jijkoun & De Rijke 2007] Jijkoun, V., & De Rijke, M. (2007, September). Overview of webclef 2007. In Workshop of the Cross-Language Evaluation Forum for European Languages (pp. 725-731). Springer, Berlin, Heidelberg.

[Jurafsky & Martin, 2008] Jurafsky, D., & Martin, J. H. (2008). Speech and Language Processing.

[Jurczyk, & Choi, 2015] T. Jurczyk and J. D. Choi. 2015. Semantic-based Graph Approach to Complex Question-Answering. Proc of NAACL-HLT 2015 Student Research Workshop (SRW). 140-146.

[Kabadjov et al., 2013] Kabadjov, M., Steinberger, J., Steinberger, R.: Multilingual statistical news summarization. In Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R., eds.: Multilingual Information Extraction and Summarization. Volume 2013 of Theory and Applications of Natural Language Processing. Springer Berlin Heidelberg (2013) 229–252

[Kalady et al., 2010] Saidalavi Kalady, Ajeesh Elikkottil, and Rajarshi Das. 2010. Natural language question generation using syntax and keywords. In The 3rd Workshop on Question Generation.

[Kanaan et al., 2009] G. Kanaan, A. Hammouri, R. Al-Shalabi and M. Swalha, "A New Question Answering System for the Arabic Language", American Journal of Applied Sciences 6 (4). 2009. 797-805.

[Katz et al., 2002] Katz, Felshin, B., Yuret, S., Ibrahim, D., Temelkuran, B., 2002. Omnibase: uniform access to heterogeneous data for question answering. In: Proc of the 7th International Workshop on Application.

[Khader et al., 2016] Khader, M., Awajan, A., & Alkouz, A. (2016). Textual Entailment for Arabic Language based on Lexical and Semantic Matching. International Journal of Computing & Information Sciences, 12(1), 67.

[Kilgarriff & Grefenstette, 2001] Kilgarriff, A., & Grefenstette, G. (2001, March). Web as corpus. In Proceedings of Corpus Linguistics 2001 (pp. 342-344). Corpus Linguistics. Readings in a Widening Discipline.

[Kolomiyets & Moens, 2011] Kolomiyets, O., & Moens, M. F. (2011). A survey on question answering technology from an information retrieval perspective. Information Sciences, 181(24), 5412-5434.

[Kor, 2005] Kor, K. W. (2005). Improving answer precision and recall of list questions.School of Informatics University of Edinburgh.

[Kurdi et al., 2014] Kurdi, H., Alkhaider, S., & Alfaif, N. (2014). Development and evaluation of a web based question answering system for arabic language. Computer Science & Information Technology (CS & IT), 4(02), 187-202.

[Kwok et al., 2001] Kwok, C., Etzioni, O., Weld, D.S., 2001. Scaling question answering to the web. ACM Trans. Inf. Syst. (TOIS) 19 (3), 242–262.

[Kwok et al., 2005] Kwok K-L., Choi S., Dinstl N. and Deng P. (2005). ―NTCIR-5 Chinese, English, Korean Cross Language Retrieval Experiments using PIRCS‖, In: Proc. of the Fifth NTCIR Workshop Meeting. NII, Tokyo, pp.88-95.

[Larkey & Connell, 2001] Larkey, L. S., & Connell, M. E. (2001). Arabic Information Retrieval at UMass in TREC-10. In TREC.

[Larson, 2009] Larson, R. R. (2009). Interactive probabilistic search for GikiCLEF. InMultilingual Information Access Evaluation I. Text Retrieval Experiments (pp. 334-341). Springer Berlin Heidelberg.

[Laurent, et al. 2010] D. Laurent, P. Séguéla et S. Nègre (2010). 'Cross lingual question answering using qristal for clef 2006'. Evaluation of Multilingual and Multi-modal Information Retrieval pp. 339–350.

[Lee et al. , 2005] Lee, Cheng-Wei, Cheng-Wei Shih, Min-Yuh Day, Tzong-Han Tsai, Tian-Jian Jiang, Chia-Wei Wu, Cheng-Lung Sung, Yu-Ren Chen, Shih-Hung Wu, and Wen-Lian Hsu. 2005. Asqa: Academia sinica question answering system for ntcir-5 clqa. In Proceedings of the 5th NTCIR Workshop Meeting (NTCIR-5), Tokyo, Japan, December.

[Lehnert, 1977] Lehnert, W.(1977). A conceptual theory of question answering, In Proceedings 5thInternational Joint Conference on Artificial Intelligence, pp.158-164.

[Lesk, 1986] Lesk, M. (1986, June). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th annual international conference on Systems documentation (pp. 24-26). ACM.

[Li & Croft, 2001] Li, X., & Croft, W. B. (2001, March). Evaluating question-answering techniques in Chinese. In Proceedings of the first international conference on Human language technology research (pp. 1-6). Association for Computational Linguistics.

[Ligozat, 2003] Ligozat, A. L. (2003). Système de Question Réponse: apport de l'analyse syntaxique lors de l'extraction de la réponse. Actes des 6e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues.

[Ligozat, 2006] Ligozat, A. L. (2006). Exploitation et fusion de connaissances locales pour la recherche d'informations précises (Doctoral dissertation, Ph. D. thesis, Université Paris-Sud 11, Orsay, France).

[Ligozat, 2013] Ligozat, A. L. (2013, August). Question Classification Transfer. In ACL (2) (pp. 429-433).

[Lopez et al., 2011] Lopez, V., Uren, V., Sabou, M., & Motta, E. (2011). Is question answering fit for the semantic web?: a survey. Semantic Web, 2(2), 125-155.

[Magnini et al., 2004] Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., De Rijke, M., ... & Sutcliffe, R. (2004, September). Overview of the CLEF 2004 multilingual question answering track. In Workshop of the Cross-Language Evaluation Forum for European Languages (pp. 371-391). Springer, Berlin, Heidelberg.

[Magnini et al., 2006] Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., ... & Sutcliffe, R. (2006, September). Overview of the CLEF 2006 multilingual question

answering track. In Workshop of the Cross-Language Evaluation Forum for European Languages (pp. 223-256). Springer, Berlin, Heidelberg.

[Manning & Jurafsky, 2012] Manning C. and Jurafsky D., "StanfordNLP Group Official Website", http://nlp.stanford.edu/software/index.shtml, checked July 14th, 2012.

[Mark et al., 2002] Greenwood, Mark, Ian Roberts, and Robert Gaizauskas. 2002. University of she_eld trec 2002 q & a system. In E. M. Voorhees and Lori P. Buckland, editors, The Eleventh Text REtrieval Conference (TREC-11), Washington. U.S. Government Printing Office. NIST Special Publication 500-XXX.

[Mendes & Véronique, 2004] Mendes, S., and Véronique M., (2004), "L'analyse des questions: intérêt pour la génération des réponses." Workshop Question-Réponse. 2004.

[Mervin, 2013] Mervin, R. An Overview of Question Answering System. International Journal of Research In Advance Technology in Engineering (IJRATE), 1, 2013

[Metzler & Croft, 2004] Metzler, D., & Croft, W. B. (2004). Analysis of Statistical Question Classification for Fact-based Questions. In Journal of Information Retrieval.

[Mihalcea & Moldovan 2001] Mihalcea, R., & Moldovan, D. I. (2001). extended wordnet: Progress report. In in Proceedings of NAACL Workshop on WordNet and Other Lexical Resources.

[Mihalcea & Radev, 2011] Mihalcea, R., & Radev, D. (2011). Graph-based natural language processing and information retrieval. Cambridge University Press.

[Mishra & Jain, 2015] Mishra, A., & Jain, S. K. (2015). A survey on question answering systems with classification. Journal of King Saud University-Computer and Information Sciences.

[Mitamura et al., 2007] Mitamura T., Lin F., Shima H., Wang M., Ko J., Betteridge J., Bilotti M., Schlaikjer A. and Nyberg E. (2007). "JAVELIN III: Cross-Lingual Question Answering from Japanese and Chinese Documents", In Proceedings of NTCIR-6 Workshop.

[Mohamed et al ., 1973] MOHAMMED, F. A., NASSER, Khaled, et HARB, H. M. A knowledge based Arabic question answering system (AQAS). ACM SIGART Bulletin, 1993, vol. 4, no 4, p. 21-30.

[Moldoan & Rus, 2001] Dan Moldovan and Vasile Rus. Logic Form Transformation of WordNet and its Applicability to Question Answering. In Proceedings of ACL 2001.

[Moldovan et al., 2000] Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Girju, R., Goodrum, R., & Rus, V. (2000, October). The structure and performance of an open-domain question answering system. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (pp. 563-570). Association for Computational Linguistics.

[Moldovan et al., 2002] Moldovan, D. I., Harabagiu, S. M., Girju, R., Morarescu, P., Lacatusu, V. F., Novischi, A., ... & Bolohan, O. (2002, November). LCC Tools for Question Answering. In TREC.

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

[Moldovan et al., 2007] Moldovan, D., Clark, C., Harabagiu, S.M., Hodges, D., (2007), COGEX: a semantically and contextually enriched logic prover for question answering, J Appl. Logic 5 (2007) 49–69.

[Mollá et al., 2000] Diego Molla,´ Rolf Schwitter, Michael Hess, and Rachel Fournier. 2000. Extrans, an answer extraction system. Traitment Automatique des Langues, 41(2):495–522.

[Mollá et al., 2006] Mollá, Diego, Menno van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In Zakerman Covendon, Lowrence and Ingrid, editors, Proceedings of the 2006 Australasian Language Technology Workshop (ALTW 2006), pages 51–58, Sancta Sophia Collage, Sydney.

[Mollá et al., 2007] Molla, D., van Zaanen, M., Cassidy, S.: Named entity recognition in question answering of ´ speech data. In Colineau, N., Dras, M., eds.: Proc. ALTW 2007. Volume 5. (2007) 57–65

[Mollá, 2003] Mollá, D., (2003), Towards semantic-based overlap measures for question answering, in: Proceedings of the First Australasian Language Technology Workshop (ALTW'03), 2003.

[Moreale & Vargas-Vera, 2004] Moreale, Emanuela and Vargas-Vera, Maria (2004). A question-answering system using argumentation. In: Monroy, Ra´ul; Arroyo-Figueroa, Gustavo; Sucar, Luis Enrique and Sossa, Humberto eds. MICAI 2004: Advances in Artificial Intelligence. Lecture Notes in Computer Science, 2972 (2004). Berlin: Springer, pp. 400–409.

[Moriceau et al., 2010] Moriceau, V., Tannier, X., & Falco, M. (2010, July). Une étude des questions "complexes" en question-réponse. In Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2010, article court), Montréal, Canada.

[Mouelhi 2008] Mouelhi, Z. (2008, March). AraSeg: un segmenteur semi-automatique des textes arabes. In JADT 2008 (pp. 867-877). Presses Universitaires de Lyon.

[Mourad, 2001] Mourad, G. (2001). Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique de citations: réalisation des applications informatiques: SegATex et CitaRE (Doctoral dissertation, Paris 4).

[Muṣṭafá et al., 2008] MUṢṬAFA M., SAYED AHMED N., DARWICH M., ABDALLAH A. (2008). Mu'jam al-Wasīṭ. Published in Bayrūt : Dār Iḥyā' al-Turāth al-'Arabī lil-Ṭibā'ah wa-al-Nashr wa-al-Tawzī.

[Nanda, 2014] Nanda, M. (2014). The Named Entity Recognizer Framework. International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN, 2349-2163.

[Nasri et al., 2016] Nasri, M., Abouenour, L., Kabbaj, A., & Bouzoubaa, K. A novel approach for semantic analysis of Arabic texts using an Arabic ontology and Conceptual Graphs. https://scholar.google.com/citations?user=vDTfO3IAAAAJ&hl=fr

[Navigli, 2009] R. Navigli, ''Word sense disambiguation: a survey''. ACM Comput Surv 41(2):1–69, 2009.

[NBdour & Gharaibeh, 2013] Bdour W. N and Gharaibeh N. K. (2013) ―Development of Yes/No Arabic Question Answering System‖, In International Journal of Artificial Intelligence & Applications (IJAIA),. Vol.4. No.1, January 2013. DOI: 10.5121/ijaia.2013.4105 51.

[Nguyen & Huong, 2008] Nguyen, Anh Kim, and Huong, Thanh Le. "Natural language interface construction using semantic grammars." PRICAI 2008: Trends in Artificial Intelligence. Springer Berlin Heidelberg, 2008. 728-739.

[Niu & Hirst, 2004] Niu, Y., & Hirst, G. (2004, July). Analysis of semantic classes in medical text for question answering. In Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains (pp. 54-61). Association for Computational Linguistics.

[Niu et al., 2003] Niu, Y., Hirst, G., McArthur, G., & Rodriguez-Gianolli, P. (2003, July). Answering clinical questions with role identification. In Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13 (pp. 73-80). Association for Computational Linguistics.

[Nyberg et al., 2002] Nyberg, E., Mitamura, T., Carbonell, J. G., Callan, J., Collins-Thompson, K., Czuba, K., ... & Ko, J. (2002). The javelin question-answering system at trec 2002. Computer Science Department, 322.

[Nyberg et al., 2003] E. Nyberg, T. Mitamura, J. Callan, J. Carbonell, R. Frederking, K. Collins-Thompson, L. Hiyakumoto, Y. Huang, C. Huttenhower, S. Judy, J. Ko, A. Kupsc, L. V. Lita, V. Pedro, D. Svoboda, and B. Vand Durme. The javelin question-answering system at trec 2003: A multi strategy approach with dynamic planning. In TREC, 2003.

[Nyberg et al., 2005] Nyberg E., Frederking R., Mitamura T., Bilotti M., Hannan K., Hiyakumoto L., Ko J., Lin F., Lita L., Pedro V. and Schlaikjer A. (2005). "JAVELIN I and II systems at TREC 2005", In Proc. of TREC'05.

[Ofoghi et al., 2006] Ofoghi B., Yearwood J., Ghosh R., « A semantic approach to boost passage retrieval effectiveness for question answering », ACSC '06 : Proceedings of the 29th Australasian Computer Science Conference, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, p. 95-101, 2006.

[Ofoghi, 2009] Bahadorreza, O. F. O. G. H. I. (2009). Enhancing Factoid Question Answering using Frame semantic-based approaches (Doctoral dissertation, Phd Thesis, Université de Ballarat, Ballarat (Au)).

[Olvera-Lobo & Gutiérrez-Artacho, 2011] Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Multilingual Question-Answering System in biomedical domain on the Web: an evaluation. In: Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., de Rijke, M. (eds.) CLEF 2011. LNCS, vol. 6941, pp. 83–88. Springer, Heidelberg (2011)

[Olvera-Lobo & Gutiérrez-Artacho, 2015] Olvera-Lobo, M. D., & Gutiérrez-Artacho, J. (2015). Question Answering Track Evaluation in TREC, CLEF and NTCIR. In New Contributions in Information Systems and Technologies (pp. 13-22). Springer International Publishing.

[Osama et al., 2011] Badawy, Osama, Mohamed Shaheen, and Abdelbaki Hamadene. 2011. ARQA: An intelligent Arabic question answering system. In Proceedings of Arabic Language Technology International Conference (ALTIC 2011), pages 1–8, Alexandria, Egypt. Bibliotheca Alexandrina.

[Ou et al., 2008] OU, Shiyan, PEKAR, Viktor, ORASAN, Constantin, et al.Development and Alignment of a Domain-Specific Ontology for Question Answering. In : LREC. 2008.

[Pakray et al. 2009] Pakray, P., Bandyopadhyay, S., & Gelbukh, A. F. (2009, November). Lexical based two-way RTE System at RTE-5. In TAC.

[Pakray et al., 2011] Pakray, P., Bhaskar, P., Banerjee, S., Pal, B. C., Bandyopadhyay, S., & Gelbukh, A. F. (2011). A Hybrid Question Answering System based on Information Retrieval and Answer Validation. In CLEF (Notebook Papers/Labs/Workshop).

[Patil et al., 2016] Patil, N. V., Patil, A. S., & Pawar, B. V. Issues and Challenges in Marathi Named Entity Recognition. International Journal on Natural Language Computing (IJNLC) Vol, 5, 15-30. 2016.

[Payne & Reader, 2006] Payne, S.J. & Reader, W.R. (2006). Constructing structure maps of multiple on-line texts. International Journal of Human Computer Studies, 64, 461-474. doi: 10.1016/j.ijhcs.2005.09.003.

[Peñas & Rodrigo, 2011] Peñas, A., Rodrigo, Á.: A Simple Measure to Assess Non-response. In: Proceedings of 49th Annual Meeting of the Association for Compu-tational Linguistics-Human Language Technologies (ACL-HLT 2011), Portland, Oregon, USA (2011)

[Peñas et al. 2006] Peñas A, Rodrigo Á, Sama V, Verdejo F (2006) Overview of the answer validation exercise 2006. In: Peters et al (2007), pp 257–264

[Peñas et al., 2007] Peñas, Anselmo, Álvaro Rodrigo, Valentín Sama, and Felisa Verdejo. 2007. "Overview of the Answer Validation Exercise 2006." In Evaluation of Multilingual and Multi-modal Information Retrieval, 257–264. Springer.

[Peñas et al., 2009] Peñas, A., Forner, P., Sutcliffe, R., Rodrigo, Á., Forăscu, C., Alegria, I., ... & Osenova, P. (2009, September). Overview of ResPubliQA 2009: question answering evaluation over European legislation. In Workshop of the Cross-Language Evaluation Forum for European Languages (pp. 174-196). Springer, Berlin, Heidelberg.

[Peñas et al., 2010] Peñas, Anselmo, Pamela Forner, Álvaro Rodrigo, Richard F. E. Sutcliffe, Corina Forascu, and Cristina Mota. 2010. "Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation." In Multilingual Information Access I. Text Retrieval Experiments.

[Peñas et al., 2012] Peñas, A., Magnini, B., Forner, P., Sutcliffe, R., Rodrigo, A., Giampiccolo, D.: Question answering at the cross-language evaluation forum 2003—2010. Lang. Resour. Eval. 46(2), 177–217 (2012)

[Perret, 2005] PERRET, Laura. Extraction automatique d'information: génération de résumé et de question-réponse. 2005. Thèse de doctorat.

[Pho, 2012] Pho, V. M. (2012). Génération de réponses pour un système de questions-réponses. In CORIA (pp. 449-454).

[Poibeau, 2003] T. Poibeau, "Extraction automatique d'information," Du texte brut au web sémantique. Hermès. 2003. P.250 pages.

[Pradhan, et al. 2005] Pradhan, S., Ward, W., Hacioglu, K., Martin, J. H., & Jurafsky, D. (2005, June). Semantic role labeling using different syntactic views. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 581-588). Association for Computational Linguistics.

[Pudaruth, et al., 2016] Pudaruth, S., Boodhoo, K., & Goolbudun, L. (2016, March). An intelligent question answering system for ICT. In Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on (pp. 2895-2899). IEEE.

[Quaero, 2008] Quaero (2008). 'Le programme Quaero'. http ://www.quaero.org/.

[Quarteroni & Moschitti, 2010] Quarteroni, S., & Moschitti, A. (2010, June). A Comprehensive Resource to Evaluate Complex Open Domain Question Answering. In LREC.

[Quintard, 2009] L. Quintard (2009). 'Overview of the QUAERO 2008 monolingual question-answering track'. http ://www.lne.eu/en/r_and_d/quaero.asp.

[Radev et al., 2002] Radev, D.R., Qi, H., Wu, H., Weiguo, F.: Evaluating Web-based Question Answering Systems. In: Proceedings LREC (2002)

[Rao et al., 2013] Rao, P. R., Devi, S. L., & Rosso, P. (2013). Automatic Identification of Concepts and Conceptual relations from Patents Using Machine Learning Methods. ICON, 2013, 18-20.

[Rastier, 2005] Rastier F. (2005). « Enjeux épistémologiques de la linguistique de corpus », in : Williams C. G. (dir), La linguistique de corpus, Rennes : P.U.R

[Razmara, 2008] M. Razmara, Answering list and other questions, PhD thesis, Concordia University, August 2008.

[Resnik, 1998] Resnik, P., (1998), Parallel strands: A preliminary investigation into mining the web for bilingual text, in conference of the association for machine translation in the Americas, 1998.

[Rinaldi et al., 2004] Rinaldi, F., Dowdall, J., Schneider, G., & Persidis, A. (2004, July). Answering questions in the genomics domain. In Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains(pp. 46-53).

[Rodrigo et al., 2010] Rodrigo, Á., Perez-Iglesias, J., Peñas, A., Garrido, G., and Araujo, L. (2010). A Question Answering System based on Information Retrieval and Validation. InCLEF (Notebook Papers/LABs/Workshops).

[Rosso et al. 2006] P. Rosso, Y. Benajiba and A. Lyhyaoui, "Towards an Arabic Question Answering system", In Proceedings of 4th Conference on Scientific Research Outlook & Technology Development in the Arab world. SROIV. Damascus, Syria. 11-14 December. 2006.

[Rosso et al., 2005] Rosso, P., Lyhyaoui, A., Peñarrubia, J., y Gómez, M. M., Benajiba, Y., & Raissouni, N. (2005, June). Arabic-English question answering. In Proc. of Information Communication Technologies Int. Symposium (ICTIS), Tetuan, Morocco, June.

[Ruppenhofer, et al. 2006] Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., & Scheffczyk, J. (2006). FrameNet II: Extended theory and practice.

[Salloum, 2009] Salloum, W. (2009, November). A question answering system based on conceptual graph formalism. In Knowledge Acquisition and Modeling, 2009. KAM'09. Second International Symposium on (Vol. 3, pp. 383-386). IEEE.

[Séjourné, 2009] K. Séjourné, Questions réponses et interactions, PhD thesis, Universite Paris Sud XI, décembre 2009.

[Sheker et al., 2016] Sheker, M., Saad, S., Abood, R., & Shakir, M. (2016). Domain-Specific Ontology-Based Approach For Arabic Question Answering. Journal of Theoretical and Applied Information Technology,83(1).

[Shima et al., 2011] Shima, H., Kanayama, H., Lee, C., Lin, C., Mitamura, T., Miyao, Y., … Takeda, K. (2011). Overview of NTCIR-9 RITE: Recognizing Inference in TExt. In NTCIR-9 Workshop (pp. 291–301). inproceedings.

[Simmons, 1965] R.F. Simmons, Answering English questions by computer: A survey, Communications of the ACM 8 (1) (1965) 53–70.

[Sinclair, 2005] Sinclair, J. (2005). Corpus and text - basic principles. In M. Wynne (Ed.), Developing linguistic corpora: A guide to good practice (pp. 1–16). Oxford, UK: Oxbow Books.

[Sing et al., 2005] Sing, G.O., Ardil, C., Wong, W., Sahib, S.: Response Quality Evaluation in Heterogeneous Question Answering System: A Black-box Approach. In: Proceedings of World Academy of Science, Lisbon, vol. 9 (2005)

[Soubbotin et al, 2002] Soubbotin, M. M., & Soubbotin, S. M. (2002, November). Use of Patterns for Detection of Likely Answer Strings: A Systematic Approach. In TREC (Vol. 52, p. 90).

[Sowa, 1984] Sowa John F. (1984). Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley Company.

[Stenchikova et al. 2006] Stenchikova, S., Hakkani-Tür, D., & Tur, G. (2006). QASR: Question answering using semantic roles for speech interface. In Ninth International Conference on Spoken Language Processing.

[Suresh kumar & Zayaraz, 2014] Suresh kumar, G., Zayaraz, G., 2014. Concept relation extraction using Naıve Bayes classifier for ontology-based question answering systems. J. King Saud Univ.

[Susan et al, 2002] Dumais, Susan, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. 2002. Web question answering: is more always better? In Proc. 25th ACM SIGIR, pages 291{298, Tampere, Finland.

[Tahri & Tibermacine, 2013] Adel Tahri and Okba Tibermacine. Dbpedia Based Factoid Question Answering System. In International Journal of Web & Semantic Technology (IJWesT) Vol.4, No.3, July 2013.

[Tari & Baral, 2005] Tari, L., Baral, C., (2005), Using AnsProlog with Link Grammar and WordNet for QA with deep reasoning, in: Proceedings of the AAAI 2005 Workshop Inference for Textual Question Answering, AAAI Press, Pittsburgh, PA, USA, 2005, pp. 13–21.

[Tatu et al. 2006] M. Tatu, B. Iles, and D. Moldovan. Automatic Answer Validation using COGEX. In Workshop CLEF 2006, Alicante, Spain, 2006.

[Teney et al., 2016] Teney, D., Liu, L., & Hengel, A. V. D. (2016). Graph-structured representations for visual question answering. arXiv preprint arXiv:1609.05600.

[Terol et al., 2007] Terol, R M., Martínez-Barco, P., Palomar, M., (2007), A knowledge based method for the medical question answering problem, Computers in Biology and Medicine, vol. 37, n° 10, p. 1511-1521, 2007.

[Toney et al., 2008] D. Toney, S. Rosset, A. Max, O. Galibert, and E. Bilinski. 2008. An Evaluation of Spoken and Textual Interaction in the RITEL Interactive Question Answering System. In Proceedings of the 6th International Language Resources and Evaluation (LREC'08), Morocco.

[Trigui et al. 2012] Omar Trigui, Lamia Hadrich Belguith, Paolo Rosso, Hichem Ben Amor and Bilel Gafsaoui. 2012. Arabic QA4MRE at CLEF 2012: Arabic Question Answering for Machine Reading Evaluation. CLEF (Online Working Notes/Labs/Workshop).

[Trigui et al.,2010] Trigui, O., Belguith, L. H., & Rosso, P. (2010). DefArabicQA: Arabic definition question answering system. In Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC, Valletta, Malta (pp. 40-45).

[Usunier et al., 2004] Usunier N., Amini M., Gallinari P. (2004) Boosting Weak Ranking Functions to Enhance Passage Retrieval for Question Answering, In IR4QA workshop of SIGIR 2004.

[Vallin et al., 2005] Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., ... & Sutcliffe, R. (2005, September). Overview of the CLEF 2005 multilingual question answering track. In Workshop of the Cross-Language Evaluation Forum for European Languages (pp. 307-331). Springer, Berlin, Heidelberg.

[Vanderwende, 2007] Vanderwende, L. (2007, March). Answering and Questioning for Machine Reading. In AAAI Spring Symposium: Machine Reading (p. 91).

[Voorhees & Weischedel, 2000] Voorhees, E., Weishedel, R., 2000. Issues, tasks and program structures to roadmap research in question & answering (Q&A). .

[Voorhees, 2001] Voorhees, E.M., 2001. The TREC question answering track. Nat. Lang. Eng. 7 (4), 361–378.

[Voorhees & Harman, 2002] Voorhees, E. M., & Harman, D. (2002, November). Overview of TREC 2002. In Trec.

[Voorhees, 2004] Voorhees, E.M., 2004. Overview of the TREC 2003 question answering Track. Twelth Text REtrieval Conference, Volume 500–255 of NIST Special Publications, Gaithersburg, MD. National Institute of Standards and Technology.

[Weiming & Hu, 2007] Weiming, W., Hu, D., Feng, M., & Wenyin, L. (2007, October). Automatic clinical question answering based on UMLS relations. In Semantics, Knowledge and Grid, Third International Conference on (pp. 495-498). IEEE.

[Witten & Frank, 1999] Ian H. Witten and Eibe Frank. 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann.

[Wong, 2004] Wong, W., 2004. Practical approach to knowledge-based question answering with natural language understanding and advanced reasoning (MSc thesis), Kolej Universiti Teknikal Kebangsaan Malaysia.

[Woods, 1973] Woods, William A. "Progress in natural language understanding: an application to lunar geology." Proceedings of the June 4-8, 1973, national computer conference and exposition. ACM, 1973.

[Wren, 2011] Wren JD. (2011). "Question answering systems in biology and medicine – the time is now", Bioinformatics, 27(14):2025–2026.

[Wyse & Piwek, 2009] Wyse, B., & Piwek, P. (2009). Generating questions from openlearn study units.

[Zhang & Lee, 2002] Zhang, D., & Lee, W. S. (2002, November). Web Based Pattern Mining and Matching Approach to Question Answering. In TREC(Vol. 2, p. 497).

[Zhang & Lee, 2003] Zhang, D. & Lee, W., 2003. A Web-based Question Answering System. Massachusetts Institute of Technology (DSpace@MIT).

[Zheng, 2002] Zheng, Z. (2002, March). AnswerBus question answering system. InProceedings of the second international conference on Human Language Technology Research (pp. 399-404). Morgan Kaufmann Publishers Inc..

[Zhiping, 2002] Zheng, Zhiping. 2002. AnswerBus question answering system. In E. M. Voorhees and Lori P. Buckland, editors, Proceeding of HLT Human Language Technology Conference (HLT 2002), San Diego, CA, March 24-27.

[Zribi et al., 2010] Zribi, I., Hammami, S. M. and Belguith, L. H., (2010), "L'apport d'une approche hybride pour la reconnaissance des entités nommées en langue arabe", In TALN'2010, Montréal, 19-23 juillet 2010 (pp. 19–23).

[Zweigenbaum et al., 2008] Zweigenbaum, P., Grau, B., Ligozat, A. L., Robba, I., Rosset, S., Tannier, X., ... & Bellot, P. (2008). Apports de la linguistique dans les systèmes de recherche d'informations précises. Revue française de linguistique appliquée, 13(1), 41-62.

ANNEXES…

Annexe A : Sorties des différents modules de système NArQAS

Dans cette annexe, nous présentons les analyses appropriées par notre système NArQAS pour un exemple de question donnée. Nous présentons également les sorties de ses différents modules. Ces sorties sont présentées en des fichiers XML.

Question :

أين يوجد مقر منظمة اليونسكو ؟

```
<?xml version="1.0" encoding="UTF-8" ?>
- <text>
  - <p>
    - <s num="1">
        <w num="1" type="O">اين</w>
        <w num="2" type="O">يوجد</w>
        <w num="3" type="O">مقر</w>
        <w num="4" type="B-ORG">منظمة</w>
        <w num="5" type="I-ORG">اليونسكو</w>
      </s>
    </p>
  </text>
```

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
- <QuestionSyntacticAnalysis>
  - <Question>
      <Tree>(ROOT (SBARQ (WHADVP (WRB اين)) (S (VP (VBN يوجد) (NP (NN مقر) (NP (NN منظمة) (NP (DTNNP اليونسكو)))))))))</Tree>
    - <Dependencies>
        <Dependency>advmod(اين-1, يوجد-2)</Dependency>
        <Dependency>root(ROOT-0, يوجد-2)</Dependency>
        <Dependency>dobj(يوجد-2, مقر-3)</Dependency>
        <Dependency>dep(مقر-3, منظمة-4)</Dependency>
        <Dependency>dep(منظمة-4, اليونسكو-5)</Dependency>
      </Dependencies>
      <Tag>اين/WRB يوجد/VBN مقر/NN منظمة/NN اليونسكو/DTNNP</Tag>
    </Question>
  </QuestionSyntacticAnalysis>
```

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
- <Question>
  - <Terms>
      <Term Order="1" Value="اين" pos="WRB" />
      <Term Order="2" Value="يوجد" pos="VBN" />
      <Term Order="3" Value="مقر" pos="NN" />
      <Term Order="4" Value="منظمة" pos="NN" />
      <Term Order="5" Value="اليونسكو" pos="DTNNP" />
    </Terms>
  </Question>
```

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
- <Question>
  - <Concepts>
      <Concept Word="اين">اين</Concept>
      <Concept Word="يوجد">وجد</Concept>
      <Concept Word="مقر">مقر</Concept>
      <Concept Word="منظمة">منظمة</Concept>
      <Concept Word="اليونسكو">اليونسكو</Concept>
    </Concepts>
  </Question>
```

Texte :

تأسست منظمة الأمم المتحدة للتربية و العلم و الثقافة ، اليونسكو، سنة 1945. يقع مقر منظمة اليونسكو في باريس في مبنى حديث واستثنائي . ولليونسكو أيضا أكثر من 50 مكتبا ميدانيا في جميع أنحاء العالم.

يوجد مقر منظمة اليونسكو بباريس فهناك المقر الرئيسي لمنظمة اليونسكو. اليونسكو هي وكالة متخصصة تتبع منظمة الأمم المتحدة, يقوم باتباعها 191 دولة.تمتلك منظمة اليونسكو حوالي خمسة برامج أساسية هما " التربية والتعليم ، العلوم الطبيعية، العلوم الإنسانية والاجتماعية، الثقافة، الاتصالات والإعلام " .

الهدف الرئيسي من انشاء منظمة اليونسكو هو المساهمة في نشر السلام والأمن بين جميع دول العالم. منظمة اليونسكو هي منظمة عالمية كلمة اليونسكو هو اختصار يعرف به المنظمة ولكن اصلها هو منظمة الأمم المتحدة للتربية والعلم والثقافة.

تدعم اليونسكو العديد من المشاريع كمحو الأمية والتدريب التقني وبرامج تأهيل وتدريب المعلمين. يوجد مقر منظمة اليونسكو في باريس بفرنسا. تنتمي منظمة اليونسكو إلى عائلة الأمم المتحدة وتشاطرها مُثلها وأهدافها.

منظمة الأمم المتحدة للتربية و العلم والثقافة أو ما يعرف اختصاراً باليونسكو، يوجد مقرها الرئيسي في باريس.جرى تدشين المبنى الرئيسي الواقع في ساحة فونتنوا بباريس والذي يحتضن مقر اليونسكو، في 3 نوفمبر/ تشرين الثاني 1958.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <text>
- <p>
  - <s num="1">
      <w num="1" type="O">تاسست</w>
      <w num="2" type="O">منظمة</w>
      <w num="3" type="O">الامم</w>
      <w num="4" type="O">المتحدة</w>
      <w num="5" type="O">للتربية</w>
      <w num="6" type="O">و</w>
      <w num="7" type="O">العلم</w>
      <w num="8" type="O">و</w>
      <w num="9" type="O">الثقافة</w>
      <w num="10" type="B-ORG">اليونسكو</w>
      <w num="11" type="O">سنة</w>
      <w num="12" type="O">1945</w>
    </s>
  - <s num="2">
      <w num="1" type="O">يقع</w>
      <w num="2" type="O">مقر</w>
      <w num="3" type="B-ORG">منظمة</w>
      <w num="4" type="I-ORG">اليونسكو</w>
      <w num="5" type="O">في</w>
      <w num="6" type="B-LOC">باريس</w>
      <w num="7" type="O">في</w>
      <w num="8" type="O">مبنى</w>
      <w num="9" type="O">حديث</w>
      <w num="10" type="O">واستثنائي</w>
    </s>
  - <s num="3">
```

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
- <AnalyseSyntaxique>
  - <Sentence nbr_sent="1">
      <Tree>(ROOT (S (VP (VBD تاسست) (NP (NP (NP (NN منظمة) (NP (DTNN الامم) (DTJJ المتحدة))) (PP (IN ل) (NP (DTNN التربية) (CC و) (DTNN العلم) (CC و) (DTNN الثقافة)))) (NP (DTNNP اليونسكو)))
      (NP (NN سنة) (NP (CD 1945))))))</Tree>
    - <Dependencies>
        <Dependency>root(ROOT-0, 1-تاسست)</Dependency>
        <Dependency>iobj(2-منظمة, 1-تاسست)</Dependency>
        <Dependency>dep(3-الامم, 2-منظمة)</Dependency>
        <Dependency>dep(4-المتحدة, 3-الامم)</Dependency>
        <Dependency>case(5-ل, 6-التربية)</Dependency>
        <Dependency>nmod(6-التربية, 2-منظمة)</Dependency>
        <Dependency>cc(7-و, 6-التربية)</Dependency>
        <Dependency>conj(8-العلم, 6-التربية)</Dependency>
        <Dependency>cc(9-و, 6-التربية)</Dependency>
        <Dependency>conj(10-الثقافة, 6-التربية)</Dependency>
        <Dependency>dep(11-اليونسكو, 2-منظمة)</Dependency>
        <Dependency>dobj(12-سنة, 1-تاسست)</Dependency>
        <Dependency>dep(13-1945, 12-سنة)</Dependency>
      </Dependencies>
      <Tag>تاسست/VBD منظمة/NN الامم/DTNN المتحدة/DTJJ ل/IN التربية/DTNN و/CC العلم/DTNN و/CC الثقافة/DTNN اليونسكو/DTNNP سنة/NN 1945/CD</Tag>
    </Sentence>
  - <Sentence nbr_sent="2">
      <Tree>(ROOT (S (VP (VBP يقع) (NP (NN مقر) (NP (NN منظمة) (NP (DTNNP اليونسكو)))) (PP (IN في) (NP (NNP باريس))) (PP (IN في) (NP (NP (NN مبنى)) (ADJP (JJ حديث) (CC و) (JJ استثنائي)))))))
      </Tree>
    - <Dependencies>
        <Dependency>root(ROOT-0, 1-يقع)</Dependency>
        <Dependency>dobj(2-مقر, 1-يقع)</Dependency>
        <Dependency>dep(3-منظمة, 2-مقر)</Dependency>
        <Dependency>dep(4-اليونسكو, 3-منظمة)</Dependency>
        <Dependency>case(5-في, 6-باريس)</Dependency>
```

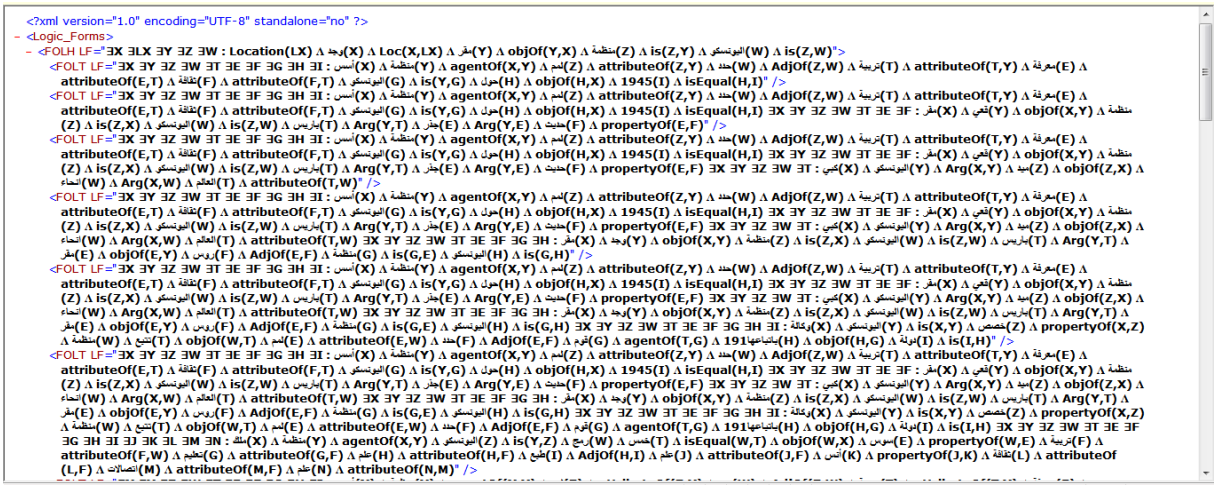
Ordinateur | Mode protégé : désactivé 100%

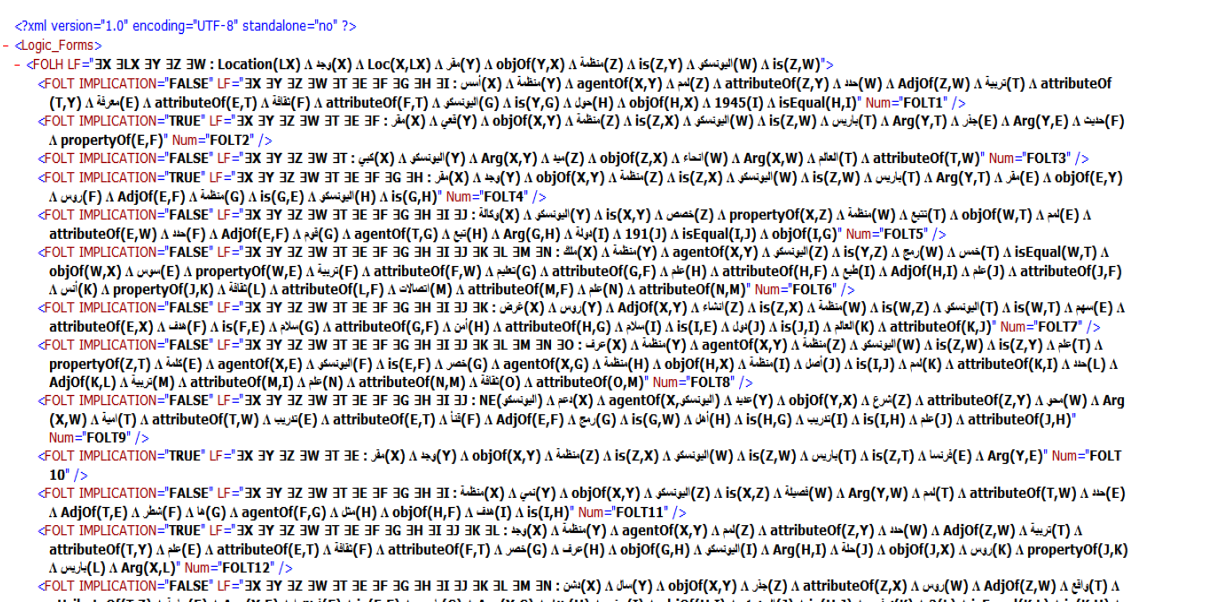
```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
- <Text>
  - <Sentence>
      <Term Order="1" Value="تاسست" pos="VBD" />
      <Term Order="2" Value="منظمة" pos="NN" />
      <Term Order="3" Value="الامم" pos="DTNN" />
      <Term Order="4" Value="المتحدة" pos="DTJJ" />
      <Term Order="6" Value="التربية" pos="DTNN" />
      <Term Order="8" Value="العلم" pos="DTNN" />
      <Term Order="10" Value="الثقافة" pos="DTNN" />
      <Term Order="11" Value="اليونسكو" pos="DTNNP" />
      <Term Order="12" Value="سنة" pos="NN" />
      <Term Order="13" Value="1945" pos="CD" />
    </Sentence>
  - <Sentence>
      <Term Order="1" Value="يقع" pos="VBP" />
      <Term Order="2" Value="مقر" pos="NN" />
      <Term Order="3" Value="منظمة" pos="NN" />
      <Term Order="4" Value="اليونسكو" pos="DTNNP" />
      <Term Order="6" Value="باريس" pos="NNP" />
      <Term Order="8" Value="مبنى" pos="NN" />
      <Term Order="9" Value="حديث" pos="JJ" />
      <Term Order="11" Value="استثنائي" pos="JJ" />
    </Sentence>
  - <Sentence>
      <Term Order="3" Value="اليونسكو" pos="DTNNP" />
      <Term Order="7" Value="50" pos="CD" />
      <Term Order="8" Value="مكتبا" pos="VBD" />
      <Term Order="9" Value="ميدانيا" pos="NN" />
      <Term Order="11" Value="جميع" pos="NOUN_QUANT" />
```

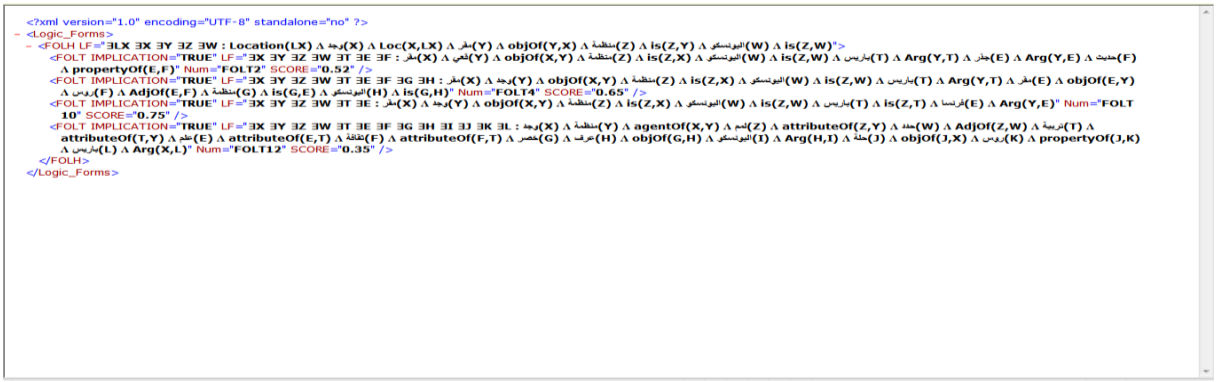
Terminé Ordinateur | Mode protégé : désactivé 100%

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
- <AllConcepts>
  - <Sentence>
      <Concept Word="تاسست">أسس</Concept>
      <Concept Word="منظمة">منظمة</Concept>
      <Concept Word="لمم">الامم</Concept>
      <Concept Word="حدة">المتحدة</Concept>
      <Concept Word="اليونسكو">اليونسكو</Concept>
      <Concept Word="1945">1945</Concept>
      <Concept Word="تربية">التربية</Concept>
      <Concept Word="معرفة">العلم</Concept>
      <Concept Word="ثقافة">الثقافة</Concept>
      <Concept Word="حول">سنة</Concept>
    </Sentence>
  - <Sentence>
      <Concept Word="يقع">قعي</Concept>
      <Concept Word="منظمة">منظمة</Concept>
      <Concept Word="اليونسكو">اليونسكو</Concept>
      <Concept Word="باريس">باريس</Concept>
      <Concept Word="ثني">استثنائي</Concept>
      <Concept Word="مقر">مقر</Concept>
      <Concept Word="جذر">مبنى</Concept>
      <Concept Word="حديث">حديث</Concept>
    </Sentence>
  - <Sentence>
      <Concept Word="اليونسكو">اليونسكو</Concept>
      <Concept Word="50">50</Concept>
      <Concept Word="كتبي">مكتبا</Concept>
      <Concept Word="ميد">ميدانيا</Concept>
      <Concept Word="اتحاء">اتحاء</Concept>
```

Terminé Ordinateur | Mode protégé : désactivé 100%







Annexe B : Règles d'Abouenour

Dans cette annexe, nous présentons les règles proposées par [Abouenour, 2014] pour extraire les relations entre concepts.

❖ Rule 1: "GTag=JJ and DTag=NN"
If the Governor(head) Tag (GTag) is "JJ" and the Dependent Tag (DTag) is a noun, then there are two cases:

- The dependent tag is neither "NNP" nor "NNPS": in this case the conceptual graph of the dendency is constructed following the pattern:

CG-dep = [cg : [Conc(G)]<-attributeOf-[Conc(D)]]

Where Conc(G) and Conc(D) are the corresponding concepts of the gouvernor and dependent respectively.

- The dependent tag is "NNP" or "NNPS": in this case the dependent is tagged by the Stanford parser as a singular or plural proper noun respectively, therefore, we follow the pattern:

CG-dep = [cg : [Conc(G) : Conc(D)]]

❖ Rule 2: "GTag ={NN,NNS} and DTag ={NNP,NNPS}"
If the GTag is NN (or plural noun NNS) and the DTag is NNP (or plural proper noun NNPS), then the pattern is:

CG-dep = [cg : [Conc(G) : Conc(D)]]

❖ Rule 3: "GTag ={NN} and DTag ={DTNN,DTNNS}"
If the GTag is NN and the DTag is DTNN or DTNNS, then the pattern is:

CG-dep = [cg : [Conc(G)]<-attributeOf-[Conc(D)]]

❖ Rule 4: "GTag ={V* } and DTag ={NN} and dependency-type=dobj"
If the GTag is a tag of a verb (such as VBP) and the DTag is NN and the dependency type returned by the Stanford parser is dobj (Direct object), then two cases occur:

- The dependent tag is neither "NNP" nor "NNPS": in this case the conceptual graph of

the dendency is constructed following the pattern:

CG-dep = [cg : [Conc(G)]<-objOf-[Conc(D)]]

- The dependent tag is "NNP" or "NNPS": in this case the dependent is tagged by the Stanford parser as a singular or plural proper noun respectively, therefore, we follow the pattern:

CG-dep = [cg : [SupConc(D) : D]<-objOf-[Conc(G)]]

Where SupConc(D)is the super concept of the NE corresponding to D.

❖ Rule 5: "GTag ={V* } and dependency-type={iobj, nsubj, dep, xcomp}"
If the GTag is a tag of a verb (such as VBP) and the dependency type returned by the Stanford parser is iobj (Indirect object), nsubj (Nominal subject), dep (General dependent) or xcomp (clausal complement with external subject) then two cases occur:

- The dependent tag is neither "NNP" nor "NNPS": in this case the conceptual graph of the dendency is constructed following the pattern:

CG-dep = [cg : [Conc(D)]<-agentOf-[Conc(G)]]

- The dependent tag is "NNP" or "NNPS": in this case the dependent is tagged by the Stanford parser as a singular or plural proper noun respectively, therefore, we follow the pattern:

CG-dep = [cg : [SupConc(D) : D]<-agentOf-[Conc(G)]]

❖ Rule 6: "GTag ={NN } and DTag ={NN}"
If the GTag is NN and the DTag is also NN, then the pattern is:

CG-dep = [cg : [Conc(G)]<-is-[Conc(D)]]

❖ Rule 7: "GTag ={CD }"
If the GTag is CD then the pattern is:

CG-dep = [cg : [Number = D]<-attributeOf-[Conc(G)]]

❖ Rule 8: "DTag ={CD }"
If the DTag is CD then the pattern is:

CG-dep = [cg : [Conc(G) = D]]

❖ Rule 9: "DTag ={JJ } and dependency-type={amod}"
If the DTag is JJ and the dependency type returned by the Stanford parser is amod (adjectival modifier), then we have two cases:

- The GTag is a tag of a verb (such as VBP): in this case no CG pattern is applied;
- The GTag is not a tag of a verb: in this case we follow the pattern:

CG-dep = [cg : [Conc(D)]<-propertyOf-[Conc(G)]]

❖ Rule 10: "dependency-type={prep}"
If the dependency type returned by the Stanford parser is a prepositional modifier (such as in, for, etc.), then the applied CG pattern is:

*CG-dep = [cg : [prep : *p i "D"]]*

Where i is the rank of the preposition D in the list of the prepositions existing in the processed text.

❖ Rule 11: "dependency-type={rcmod} and DTag={V*}"
If the dependency type returned by the Stanford parser is rcmod (Relative clause modifier), then the applied CG pattern is:

CG-dep = [cg : [Conc(G)]-attributeOf->[cg : Conc(D)]]