Arabic Question Answering from diverse data sources

الرد على الأسئلة العربية من مصادر بيانات متنوعة

by

FERAS KHATER

Dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATICS
(KNOWLEDGE AND DATA MANAGEMENT)

at

The British University in Dubai

July 2018

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

_____

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

Abstract

Currently, Arabic users are still forced to extract manually the accurate answers of their questions, which is a difficult task with a vast amount of information available on the Internet. Actually, the existing Arabic Question Answering (QA) systems do not meet the users' needs in terms of performance and scope that cover all types of questions. The motivation behind this research is the need for new approaches to handle all types of questions and answer them beyond the factoid questions. Therefore, we present in this paper a new design of the linguistic approach to develop a reliable Arabic QA system and data source with the ability to address the following challenges: (i) handle both factoid and complex questions in Arabic language, (ii) extract the precise answer from available resources, (iii) evaluate the proposed QA system based on a gold standard data set, and (iv) provide an Arabic Corpus of Occupations (ACO) corpus that has been made freely and publicly available for research purposes. Our QA system is a web application that helps us to get an answer to the question posed from different data sources. Accordingly, we conducted experiments on a set of 230 question from the previously published resources, TREC, CLEF, and Arabic Corpus of Occupations (ACO) corpus. The system performance shows an average precision of 36%, by answering 72 questions, as well as the Recall was 78% and F-Measure was 51%.

Besides, the aim that attracted us to build the Arabic Corpus of Occupations (ACO) corpus was the lack of free, annotated and large-scale Arabic resources that can be used in training and testing Arabic QA systems. In this paper, we provide ACO corpus of one million words written in Modern Standard Arabic (MSA). The corpus contains 700 occupations which are analyzed carefully and manually annotated. We use Cohen's Kappa coefficient method to evaluate the reliability of the tagged content. The corpus content has been tagged and assessed by two different groups of taggers. Accordingly, the inter-annotator agreement indicates that the reliability of ACO corpus is almost perfect agreement. As well as, the content of the corpus is highly confidence and reliable according to the result achieved by 90%.

Keywords: Arabic Language; Arabic Corpus; Part of Speech Tagging; Question Answering; Questions Answering System; Natural Language Processing.

ملخص

حاليا، لا يزال المستخدمون العرب مضطرين لاستخراج الإجابات الدقيقة لأسئلتهم بشكل يدوي، والتي تعتبر مهمة صعبة مع وجود كم هائل من المعلومات المتاحة على الإنترنت. في الواقع، لا تلبي أنظمة الرد على الأسئلة باللغة العربية احتياجات المستخدمين من حيث الأداء والنطاق الذي يغطي جميع أنواع الأسئلة. الدافع وراء هذا البحث هو الحاجة إلى أساليب جديدة للتعامل مع جميع أنواع الأسئلة والإجابة عليها. لذلك، نقدم في هذه الورقة تصميمًا جديدًا للنهج المتبع في تطوير نظام الرد على الأسئلة العربية يكون موثوق به مع القدرة على التعامل مع التحديات التالية: (1) التعامل مع الأسئلة المعقدة والبسيطة باللغة العربية، (2) استخلاص الإجابة الدقيقة من الموارد المتاحة، (3) تقييم النظام على أساس مجموعة بيانات ذات معايير معتمدة، (4) توفير مجموعة بيانات للمهن باللغة العربية متاحة للجميع لاغراض البحث. نظام الرد على الأسئلة العربية الخاص بنا هو تطبيق ويب يساعدنا في الحصول على إجابة للسؤال المطروح من مصادر بيانات مختلفة. وبناءً على ذلك، أجرينا تجارب على مجموعة من الأسئلة المكونة من 230 سؤالًا من مجموعات مختلفة مثل TREC ، و CLEF ، وACO. اظهر اداء النظام ان متوسط الدقة هي 36٪ ، من خلال الإجابة على 72، وكذلك متوسط الاستدعاء هو 78٪ و F-Measure كان 51٪.

إلى جانب ذلك، كان الهدف الذي جذبنا إلى بناء مجموعة بيانات للمهن العربية هو الافتقار إلى الموارد العربية المجانية والمشروحة والواسعة النطاق والتي يمكن استخدامها في تدريب واختبار أنظمة الرد على الاسئلة العربية. في هذه الورقة، نوفر مجموعة بيانات مكونة من مليون كلمة مكتوبة باللغة العربية الفصحى الحديثة. تحتوي المجموعة على 700 مهنة تم تحليلها بعناية ووضع علامات عليها يدويًا. تم استخدام طريقة معامل كوهين كابا لتقييم موثوقية محتوى هذه المجموعة. ايضا تم وضع علامات على محتوى المجموعة وتقييمها بواسطة مجموعتين مختلفتين من واضعي العلامات. وبناءً على ذلك، يشير اتفاق Inter-annotator إلى أن موثوقية ACO corpus هي اتفاق شبه كامل. وعليه، فإن محتوى المجموعة يتميز بدرجة عالية من الثقة والموثوقية وفقًا للنتيجة التي تم تحقيقها بنسبة 90٪.

Dedication

First and foremost, I dedicate my dissertation work to my beloved parents for their prayers, unlimited support, and encouragement that has continued throughout my life. Moreover, to my Wife, for her relentless care and support over the years. As well as, my children "Zain", "Abdulrahman", and "Zaha" for their patience during the completion of this thesis.

Acknowledgement

At the outset, all my prayers and thankfulness are to Allah the almighty for facilitating this work and giving me the strength and the ability to complete this thesis. This dissertation would not be accomplished without the help and cooperation of many people whom I would like to thank.

I would like to start by acknowledging my supervisor Prof. Khaled Shaalan for guidance, valuable support, and motivation during my study. He was always available with an accurate advice, or an interesting suggestion and a listening ear. I have been greatly appreciated.

Many thanks to those who contributed to this dissertation either directly or indirectly. I am also thankful to all my friends in UAE and Jordan.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviations	**Description**
NLP	Natural Language Processing
MT	Machine Translation
PoS	Part of Speech
IR	Information Retrieval
SEs	Search Engines
MSA	Modern Standard Arabic
QA	Question Answering
NER	Named Entity Recognition
WSD	Word Sense Disambiguation
AI	Artificial Intelligence
TREC	Text REtrieval Conference
CLEF	Cross-Language Evaluation Forum
BAMA	Buckwalter Arabic Morphological Analyzer

Chapter One: Introduction

This chapter introduces a background about QA concept in general and its problems. In addition, it presents the motivations, aim and objectives of this paper. Research questions are also proposed along with dissertation structure.

1.1. Background

Historically, Arabic is one of the most important languages in the world that has a religious specificity in Islam. It is an important language due to the Holy Quran, which was revealed and written in Arabic. Also, there are more than 1.6 billion Muslims who read the Quran in Arabic. In ancient times, Arabic language had the main role in science such as mathematics, medicine, chemistry, etc. Geographically, there are around 25 countries in the world speaking Arabic, which makes it the fifth most common language in the world (Shaalan 2010).

Accordingly, researchers were encouraged in various fields of Arabic research, particularly in Natural Language Processing (NLP). NLP helps Arabic users in their countries through Search Engines (SEs), social media, emotional analysis, etc. In addition, it assists non-Arabic speakers through automated Machine Translation (MT). Some NLP tasks have reached a high level of maturity such as Part-of-Speech (PoS) tagging, morphological analysis and stemming, which facilitated the development of Arabic NLP systems (Albarghothi, Khater & Shaalan 2017).

Due to influence of the technological revolution, the Internet has become a major role in people's lives and corporate strategies that develop Web content (digital libraries, newspapers collections, etc.) (Salloum et al. 2018). Consequently, demand for the development of more sophisticated systems has increased. Therefore, Arabic NLP has gained importance as a solution to fulfil user's requirements in terms of text translation, information retrieval, etc.

Indeed, Information Retrieval (IR) systems provide solutions for Arabic users to find relevant content and/or extract the right answers to their questions, which go beyond the current SEs used. Unfortunately, the development of Arabic IR faces two main challenges: (i) immaturity of useful tasks such as query and answer analyzer; (ii) availability and usability of Arabic NLP

resources such as corpora, lexicons, ontologies, knowledge bases, etc. (Shaalan, Al-Sheikh & Oroumchian 2012).

Hence, researchers still have enormous opportunities to develop Arabic IR systems. It is logical to work on the case of Arabic QA systems that fall within the category of Advance IR system. The Arabic QA system can provide valuable assistance to users in exploiting the Arabic web content or any set of documents. Unlike the available SEs, which provide a list of documents for user's questions. The main purpose of these systems is to enable the computer to provide accurate answers directly to user's questions that expressed in natural language. In order to process question automatically, the QA system requires the basic NLP tasks such as Phrase Chunking and PoS tagging, as well as more complex tasks such as Query and Answer Analysis, Named Entity Recognition (NER), syntactic parsing, etc. (Al-Chalabi, Ray & Shaalan 2015).

1.2. Problem Description

Currently, the Arabic users are still forced to extract manually the accurate answers of their questions, which is a difficult task given the large amount of web information. Actually, the existing Arabic QA systems do not meet the users' needs in terms of performance and scope that cover all types of questions (Cheddadi 2014). In order to fill these gaps and rely on the analysis of previous experiences, the most important challenges that could be faced in the development of the Arabic QA system were described as follows:

- **Language challenge**: in line with the works that have been done in Arabic QA systems. The main issue is how to handle the language specificities at various levels of processing questions and answers. Arabic language is morphological, derivational and inflectional language. As well as other characteristics such as no capital letters for recognizing name entities, no requirements of diacritics ambiguity and flexibility in transcribing foreign names. These challenges are considered one of the real challenges for NLP tasks (PoS tagging, NER, etc).
- **Web source challenge**: as part of the question processing, QA systems work to elicit the answer of the question by looking at a set of documents. In addition, the vast amount of information available on the Internet and the excessive interest of Arabic users in QA systems have encouraged researchers in NLP to target the Web as a source of information. Currently,

such systems are not available for Arabic language, whereas the existing systems are document retrieval that do not serve the goal of QA systems.

- **Question types challenge**: this depends on the complexity of the questions and their types as well as the expected answers. The range of these types begins with factoid and definitional questions where the expected answer is a NE, such as When, Where, Who and What. On the other hand, the List, HOW and WHY questions need more advanced processing where the expected answer is more complex. The current issue in most Arabic QA systems is that the scope of work is limited to factoid questions, and overall performance is still unsatisfactory compared to other languages. Moreover, there is no clear and understandable mechanism for handling complex questions and extracting answers from different sources.
- **Evaluation campaigns challenge**: these campaigns help researchers to assess and benchmark the performance of their systems according to standard measurements and golden dataset, which is considered a success factor in QA field. The succession of evaluation campaigns for QA systems such as CLEF and TREC have a positive impact on improving system's performance, as well as developing more advanced QA tasks. However, the Arabic language has remained absent in most of these campaigns, which led to emergence many obstacles in the evaluation process of Arabic QA systems. With regard to a golden dataset, few resources are available for evaluation such as TREC 2001 and TREC 2002, as well as the absence of a large dataset covering question types, length, domain, etc.

1.3. Research Question

In view of the above mentioned regarding the Arabic QA challenges, it is worth to ask the following questions which we will try to answer them in this thesis:

- Is it possible to build a QA system that can answer different types of Arabic questions (Factoid, Definitional and Complex)?
- Is it possible to achieve acceptable performance even with different data source?
- Is it possible to build an Arabic annotated corpus to be used as one of data source?

1.4. Research Objectives

The objectives of this study are prepared to answer the research questions, as well as to address the challenges mentioned in the above section. In line with this, these main objectives are summarized as follows:

- In the language challenge, all stages of the QA system will be studied and analyzed to find out new resources and NLP tools. A new approach will be proposed to improve the main modules of QA system, such as retrieving and ranking passages considering the Arabic language characteristics. In addition, investigate the effectiveness of current Arabic resources and tools for better integration such as stemming, PoS tagging, NER, etc.
- According to previous research conducted in this field, we found that the most important module in the QA architecture pipeline is the passage retrieval. Therefore, focus on providing high-quality passages that belong to the question in order to improve the performance of QA systems at all.
- In the Web challenge, introduce and design a mechanism that enables Arabic users to get answers for their questions from the different resources.
- In the question types challenge, all types of questions will be studied and analyzed to predict the correct answers in order to meet the needs of users, as well as examples of Arabic questions will be analyzed to determine the causes of failure in previous approaches.
- In the evaluation challenge, QA experiments will be conducted using a relevant large set of questions and well-known measurements based on golden dataset such as TREC and CLEF. Consequently, performance results will be compared with other QA systems.

1.5. Dissertation Structure

This dissertation consists of several chapters; the current chapter is the introductory chapter that provides a background about QA concept in general and its problems, as well as it presents the motives, objectives, and research questions of this thesis. The remaining chapters are organized as follows:

- Chapter two presents an extensive study of the existing works in the Arabic QA field and the related NLP tasks, as well as the information resources. It highlights the earlier approaches

and QA systems to reveal key issues and ways in which these issues were addressed. In addition, it demonstrates the corpora roles and the available Arabic information resources, as well as the evaluation campaigns in various NLP tasks.

- Chapter three introduces the methodology of building Arabic corpus, so that some positive progress related to this subject can be added. This provides a large-scale Arabic annotated resource, as well as mature content written in MSA format selected from appropriate sources.
- Chapter four provides the system framework where we developed a new QA system that handle all types of questions beyond the factoid questions. This tool utilizes existing NLP tasks and available information resources.
- Chapter five draws the overall conclusions for the key issues that are addressed in this dissertation. In addition, discuss the answers to research questions, as well as provide the future recommendation.

Chapter Two: Literature Review

This chapter illustrates the major components of QA system and focuses on the challenges of Arabic QA to identify the characteristics that help in the production of such systems for the Arabic language. It also highlights the existing systems of Arabic QA that will help us compare and measure our contribution to other systems. In addition, it demonstrates the corpora roles and the available Arabic information resources, as well as the evaluation campaigns in various NLP tasks.

2.1. General Overview

In general, the QA system needs to understand dialect language in order to answer the user's question through NLP and common knowledge. Hence, many researchers used Artificial Intelligence (AI) techniques that integrate NLP tasks and knowledge to formulate QA logics. The knowledge composes in the form of rules, frames, logic, templates, and ontology that used to process of matching the questions and answers. Linguistic techniques such as tokenization, PoS tagging and parsing are utilized to reformulate the user's question into a special query that extracts the answer from a knowledge base or corpus. Nevertheless, building an appropriate knowledge base is a time-consuming process, since QA systems rely on huge information for a particular domain which needs a long time to produce it. In addition, publishing a specific domain knowledge base is not effective because the different system will require different knowledge base and rules.

Attention has begun to design QA systems in the 1960s with the attempt by Green et al (1961) to produce a system based on structured database query called BASEBALL and LUNAR to answer English questions. The QA systems that implemented early focused on specific domains. Moreover, the questions used by these systems were analyzed using NLP techniques in order to create a standard query for the database based on the predefined model. The information stored in the structured database was only able to answer the questions raised in the restricted domain, so this was the main limitation of these systems (Woods 1973).

Recently, the knowledge base is no longer the only source of information, as new resources are becoming available such as Internet information. A new approach has been introduced to exploit online information to answer the user's question, which allows users to extract an answer

to an unexpected question, not only a predefined question (Clark, Thompson & Porter 1999). Morphological analysis techniques are the basis of most NLP systems in order to process and answer different types of user's questions. System efficiency is measured through an assessment process that uses test sets relevant to target languages (Al-Sughaiyer & Al-Kharashi 2004).

2.2. Question Answering and Search Engine

Technically, QA systems are quite different from SEs, such as Google and Yahoo. These systems attempt to extract and display the precise answer to the user's questions without displaying lists of candidate web links and documents for further manual searching (Kwok, Etzioni & Weld 2001). While both SEs and QA systems enable users to submit questions in natural language but each one has a different technique. Undoubtedly, traditional SEs help users who are looking for information by rephrasing the user's question into keywords or logical expressions, and then provide comprehensive results in the form of documents lists. In contrast, QA systems are more convenient for users who are looking for a precise and direct answers to their questions without having to browse through a list of documents (Cheddadi 2014).

In QA systems, the user's question is processed to elicit the question's features and classify it by type and/or domain, while the targeted data source is processed to extract precise answer that often consist of documents, web pages, etc. The process of identifying the type of question (e.g. Factoid, List, Definition, etc.) is an important step in determining the type of answer. For instance, "*Who is the Prime Minister of Germany?*", the type of expected answer is "Person" and question's type is "Factoid", which means that system looks only for person names in the data source without parsing all sentences. Table 1 compares the types of questions that are utilized in QA systems to identify the type of answer. Therefore, the question type plays a key role in QA systems realization, as well as it considers one of the main challenges facing these systems. Obviously, when the scope of the QA systems handles only the factoid questions, the processing of question and answer may not require advanced techniques and tasks. Whereas the QA systems handle all types of questions, the processing of question and answer require advanced techniques for a deep understanding and learning (Mishra & Jain 2016).

Type	Expected Answer	Example	Characteristics
Factoid	NE (Person, Organization, Location, etc.)	Who is the vice president of the USA? ***Answer:*** Mike Pence	• Easy answer extraction due to the answer consists of one or two NE • Answer evaluation is easy as well
Definition	Information about a NE	Who is Sheikh Zayed? ***Answer:*** he is the Ruler of Abu Dhabi from 1922-1926. He was the youngest of Sheikh Sultan's four sons.	• Not easy answer extraction due to the answer consists of sentences and can be collected from different paragraphs • Evaluation of the answer is difficult
List	List of NEs	What are the most visited places in Dubai? ***Answer***: Burj Khalifa, Dubai Mall, Dubai Museum, Burj al-Arab, Jumeirah Beach	• Not easy answer extraction due to the expected list is scattered over various documents and paragraphs • Evaluation of the answer is difficult
Other	Yes/No, Facts, Arguments, etc.	Does UAE belong to the Arab Nations League? ***Answer:*** YES	• Not easy answer extraction due to inference, machine reading, etc. tasks are required • Evaluation of the answer is difficult, except YES/NO questions

Table 1. *Questions types*

Mishra & Jain (2016) stated another factor that plays a role in QA systems realization, which is the domain of the questions. The implementation of a close-domain (e.g. Movies, Sports, Education, etc.) for QA systems helps to reduce the challenges of question type and language. This is appropriate for users looking for specific answers in a particular domain, which means that answers are limited due to a specific repository. In contrast, open-domain system might be related to any question subject. These systems required to process a large amount of textual data related to the potential answer.

2.3. Question Answering Component

The QA system architecture consists of three major modules: Question Analysis (QA) Module, Passage Retrieval (PR) Module, and Answer Extraction (AE) Module. Figure 1 shows a pipeline of three main modules of the QA system.

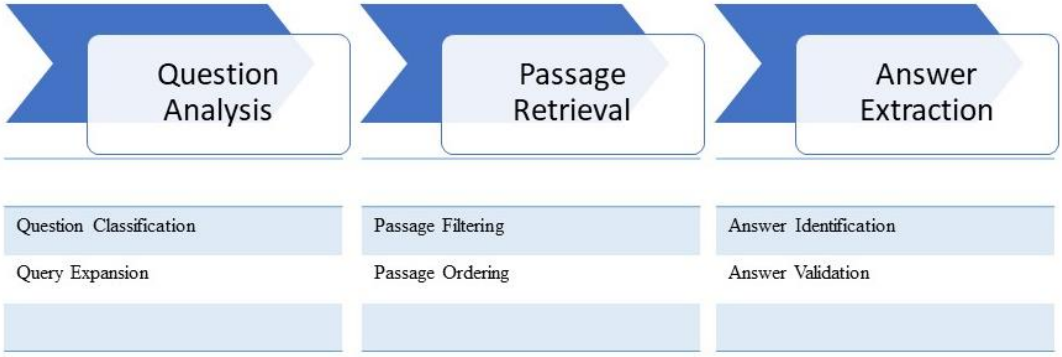

Figure 1. *General QA architecture*

2.3.1. Question Analysis

The aim of this module is to understand the question meaning and purposes. The question in natural language is analyzed to identify its type and extract its pattern, as well as structure the form of expected answer. Moreover, define the constraints of the expected answer and reformulate the query to be passed to the PR module. Initially, a morpho-syntactic analysis of question's words is performed and PoS is added to each word to indicate whether the word is the verb, singular noun, plural noun, etc. This makes it easier to know the type of information presented by the question and its purpose. Morphological analysis is utilized to identify the class of question which is essential to extract the precise answer by forcing the answer extraction module to select and validate the reasonable answers. The question classification determines the probable classes from the keywords of question, such as Name, Date, Location. For example, the system analyzes the question "Who is the vice president of the USA?" and recognize the answer is person's name, which lead to reduce the search area and looking at names only (Al Chalabi, Ray & Shaalan 2015).

In order to percept the Arabic question, it needs to be addressed by NLP tasks through returning words to their roots, because the Arabic words were built of three or four roots of characters. The words derivation is formed by joining the affixes (infix, prefix, suffix) to each root (Shaalan & Raza 2009).

Practically, all QA systems comprise this module, which is worthy to focus on its performance and functioning due to its impact on the whole performance of the QA system. However, question classification helps the system to specify the boundaries of an answer, as well

as the terms in the question help in matching and ordering the sentences of candidate answer (Rahman 2015). There are many techniques used to classify a question, one of them pattern matching which is a simple and effective technique. In addition to another technique known as heuristic rule-based algorithm, which requires an effort to define the rules (Bronner & Monz 2012).

Rosso et al. (2005) have analyzed the Arabic questions by eliminating the stop words, utilizing the NEs, and characterizing the questions in four classes: Name, Date, Quantity, and Definition. Thereafter, studies continued to develop this module by analyzing patterns of Arabic question and then identifying appropriate names and lexical items in order to improve the process of formulating questions and answers. This study used a small sample size (100 questions) which make it difficult to reach general conclusions (Kanaan et al. 2009).

2.3.2. Passage Retrieval

Typically, IR systems such as Google or Yahoo are utilized to restore the documents and passages with ranking. Therefore, the user needs to describe the query precisely via the search engine to obtain relevant documents (Rahman 2015). This module is the core unit of the QA systems, which prepares a collection of required paragraphs that match the user's question. The reason for shortening documents to paragraphs is to make the systems faster as well as filter these paragraphs and rank them to improve the quality of the answer (Jurafsky & Martin 2014).

The module role is not to find out the questions' answers but to identify the documents and paragraphs that have the potential answer. In order to identify the relevant information precisely, the documents are split into passages and handle them as documents. In the end, the module identifies and ranks the required passages and passes them to the next module (AE). The technique that relies on the passage retrieval is better than the document retrieval due to the performance and quality of the processing (Navarro, Puglisi & Valenzuela 2015).

However, NLP tasks are used to choose and assess the quality of the selected documents. There are two different techniques followed in QA systems: shallow technique and deep technique. The shallow technique utilizes the keyword to find the passages and sentences by combining the similarities between the question and the type of expected answer. This technique is suitable for factoid questions which looking for facts or simple answer (Prager 2007). While the deep technique is more complex which includes many processes, such as question and answer type

analysis, question extension, answer analysis. Moreover, various NLP tasks are used such as PoS, NER, tagging, and chunk parsing (Corston et al. 2005).

Benajiba & Rosso (2007) have ranked the retrieved passages based on the question's keywords. Accordingly, assign the higher rank to passages that have a smaller distance between those keywords and passages content. Furthermore, Abouenour (2011) has applied a three-level approach based on question keywords with morphology concepts, n-gram structure, and encoded semantics in order to rank the passages. As well, Kanaan et al. (2009) have used the Cosine method to find similarities between question keywords and documents content and then rank the document accordingly.

2.3.3. Answer Extraction

This module extracts the answer from the candidate passages and validates the answer to ensure the correct answer is retrieved. Obviously, the extraction of an answer will fail if the candidate passages are irrelevant. The module obtains the expected answer and ranked passages from previous modules in order to analyze them and extract answer (Toba et al. 2014).

Most of studies have identified their own techniques, patterns, and rules to extract the answer, as well as utilize n-grams method to find semantic similarities between the question and answers (Trigui, Belguith & Rosso 2010).

Moldovan et al. (1999) have adapted pattern approach. Once the potential answer is found from the passages, the answer validation is calculated by scoring the words overlap between the question and the answers found. This score is followed to validate and rank overall user answers. Harabagiu et al. (2001) have improved the approach that used by Moldovan and others by adding Machine Learning (ML) algorithm, which enhances the scoring function. On the other hand, Srihari & Li (1999) have used a different approach. They have relied on the constraints of the question more than the answer type in choosing and ranking sentences that contain potential answers. They used such functions to rank the sentences like the number of unique words in the sentence matched and the keywords orders in the sentence compared with their order in question.

Ittycheriah et al. (2000) have produced a ranking function based on both the type of answer and the matching words. In addition, the frequency of the user's answer is measured as a criterion

for selecting answers. This statistical method represents the number of events associated with the question (Clarke et al. 2002).

2.3.4. QA Example

Based on the above architecture and for further understanding, let's simulate an example that explains the processing mechanism of each module and the output of data format. QA system modules will address the example of the input question "*What are the most places visited in Dubai in last decade*?" as follows:

- Question Analysis module identifies the question type which is "List" and the expected answer which is a list of items as well. Sometimes, the module contains predefined question and answer pattern to match the posted questions with these patterns. Suppose the question in the example matches with the "*What is X in Y [in Z]?*" pattern. Accordingly, the module extracts the form of the answer "*X in Y [in Z] {are, is}*" which belongs to the question pattern. The question pattern and the answer form will be used in the Passage Retrieval module to focus on paragraphs that contain the same structure. In addition, the module recognizes the NEs (in the example LOCATION =" Dubai"), as well as identifies the constraints (in the example "last decade") in order to filter the candidate paragraphs according to these features.
- Passage Retrieval module attempts to extract the best paragraphs from the information source (documents collection) based on the best matches with the question's features in terms of keywords and patterns. Based on the proposed example, let's suppose that such passages are extracted from documents collection as follows:

  Passage 0: "the most visited places in Dubai by Arab in the last decade are not the same for those coming from Europe"

  Passage 1: "the most visited places in Dubai by French in the last decade are Burj Al Arab, Jumeirah and Dubai Mall"

  Passage 2: "Dubai Mall is one of the most visited places in Dubai by French tourists in the last decade"

  Passage 3: "In the last decade, French tourists have most visited Burj Al Arab, Jumeirah and Dubai Mal"

- Answer Extraction module consists of two sub-modules which are concerned with answer identification and answer validation respectively. The first sub-module processes the passages that come from PR module to list the sub-contents with their features. The second sub-module validates the list of candidate answers based on their features. For instance, the "Passage 0" in the above part contains similar question keywords and pattern. The answer extraction sub-module processes the passage and return sub-content is "*not the same for those coming from Europe*". Based on this, the system selects "Europe" as a candidate answer since the expected answer is a list of places and "Europe" is tagged with the NE feature "LOCATION". However, the answer validation sub-module will reject the selected answer if additional information is available such as Europe is not a place in Dubai.

Actually, passages extraction in PR module is often more complicated and challenging due to: (i) the correct passages do not contain in documents collection; (ii) documents collection has the passages with a different structure (For instance, Passage 3 above). It is worth noting that passages can be extracted in two main approaches: (a) retrieve the relevant documents and then extract the paragraphs from them; (b) index each passage as a single document and retrieve it accordingly (Khalid & Verberne 2008).

The performance of each module is affected by the performance of the previous module in the QA pipeline. In order to build QA system has a high efficiency, it should evaluate the impact of each module in terms of NLP tasks and techniques. Moldovan et al. (2003) have mentioned a high percentage of errors in QA systems due to mistakes in the classification of questions, which is estimated at about 36%. This affects the performance of the rest of the modules. LLopis, Vicedo & Ferrández (2002) have pointed out that the quality of answers in QA systems depends on the quality of the PR module.

2.4. Question Answering Evaluation

2.4.1. Evaluation Forums

An important trend or concept emerged in 1987 related to NLP field, which focuses on evaluation campaigns in various tasks, such as text processing (Harman 1992) and speech processing (Pallett 2003). The IR researchers also continued to follow this trend which organized

by Text REtrieval Conference (TREC) in USA (Ray & Shaalan 2016). Indeed, the number of IR evaluation campaigns and the number of participants indicates the importance of these campaigns to IR researchers. Regardless of the mechanism applied to evaluate IR tasks, these campaigns have positive impact in urging researchers to spend more effort in development of NLP field (Voorhees & Harman 2005).

Since 1994, NLP and IR research communities in Europe have begun a series of ongoing evaluation campaigns for a variety of tasks, such as lexical semantics (EDMONDS & KILGARRIFF 2002), German morphological analyzers (Hauser 1994) and French PoS taggers (Grace) (Adda et al. 1998). Furthermore, Cross-Language Evaluation Forum (CLEF) has promoted research in the area of multilingual information. It focuses on testing IR systems and creating golden datasets for researchers to help them develop and evaluate their systems (Ray & Shaalan 2016). Arabic language was introduced in CLEF (2012) as a golden dataset for QA systems, which allows Arabic QA systems to evaluate and compare their results based on this addition (Agosti et al. 2007).

Interest in NLP campaigns has shown the importance of evaluation in NLP projects, and accordingly new evaluation terms and rules have been introduced and utilized. For instance, (i) progress evaluation which assesses the current state against the desired state, (ii) adequacy evaluation which assesses the proposed tool adequacy for some intended utilize, (iii) diagnostic evaluation which assesses the proposed tool to discover where and why fails, (iv) hypothesis versus reference data which assesses the proposed tool by comparing the results produced by the tool against the data created to represent the gold-standard (Mitkov 2005).

The evaluation campaigns followed a typical four-stage model as shown in Figure 2. In phases I and II, participants are permitted to adapt their tools according to final test conditions (e.g. input and output test formats) and made any adjustment required in terms of tool function or evaluation protocol. The last two phases are the actual competition where participants utilize the gold-standard data through their tools to figure out the results and send these results for adjudication in order to obtain outcomes and ranking. In general, a workshop is organized to publish the outcomes, methods that have been evaluated, and conduct discussions among participants (Mitkov 2005).

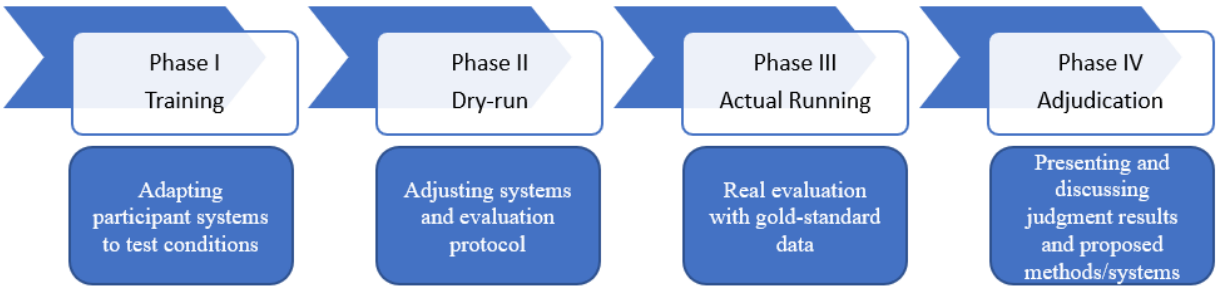

Figure 2. *Evaluation campaigns model*

The evaluation process of QA systems is carried out as a task under the IR campaigns, followed by TREC for English, CLEF for European languages, Russian Information Retrieval Evaluation Seminar (ROMIP3) for Russian language, etc. The Arabic language is one of the few languages that have one edition of evaluation for QA systems. The only editions of this language were produced in TREC (2001-2002) and CLEF (2012).

2.4.2. Performance Measures

Generally, there is a range of tests required to measure the effectiveness and performance of IR systems. These tests consist of documents collection (indexed documents according to the system needs), questions in a natural language, and a set of relevant judgments of questions. In terms of systems evaluation, two types of cases can occur, namely: ranked and unranked retrieval. In the case of rank retrieval, the evaluation depends on the list of retrieved documents and sorts them for the best match with the question pattern and keyword. While the case of unranked retrieval, the QA system is evaluated by using the following formulas (Ray & Shaalan 2016).

- Precision (P), which represents the percentage of retrieved documents related to the query

  $$P = \frac{Number\ of\ relevant\ documents}{\text{Retrieved documents}}$$

- Recall (R), which represents the percentage of related documents retrieved

  $$R = \frac{Number\ of\ relevant\ documents}{\text{Relevant documents}}$$

- F - Measure (F), which represents the percentage of combination for precision and recall

  $$F = \frac{2*P*R}{\text{P + R}}$$

The QA system can be evaluated either the whole system and / or single module. Therefore, the above formulas are not valuable because they assess the system's effectiveness in retrieving

relevant documents without highlighting the system's ability to deliver documents ranked as relevant. Various IR evaluation campaigns use more appropriate formulas to assess the quality of the overall QA systems as follows:

- Accuracy (A): the outcome of this formula is a number between 0 and 1 which indicates the probability of providing a correct answer by the QA system

  $$A = \frac{Number\ of\ correct\ answers}{\text{Number of questions}}$$

- Mean Reciprocal Rank (MRR): the outcome of this formula is a number between 0 and 1 which indicates the quality of the categorized list of possible answers (N = number of questions, rank(i) first position of relevant document)

  $$MRR = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{rank(i)}$$

- Answered Questions (AQ): the outcome of this formula is a number between 0 and 1 which indicates the probability of providing a correct answer by the QA system in the categorized list of possible answers (n = number of answered questions, |Q| = number of questions)

  $$AQ = \frac{n}{|Q|}$$

- c@1: the outcome of this formula is a number between 0 and 1 which has been used by CLEF QA tracks since 2009. In this formula, systems are encouraged to reduce the number of incorrect answers while keeping the number of correct answers and leaving some questions unanswered (nr = number of correct answered, nu = number of incorrect answered, n = number of questions)

  $$c@1 = \frac{\left(nr + nu * (\frac{nr}{n})\right)}{n}$$

2.5. Arabic Language Challenges

The Arabic language is a familiar and widespread language with nearly 300 million native speakers. It has a special context (right-to-left writing), as well as it consists of 28 letters (3 long vowels and the remaining characters are consonants). In addition, diacritics and allographic variants are utilized as short vowels except one is utilized as a double consonant marker. The

numbers are written from left to right, making it difficult for Arabic-language editors to deal with written words from left to right and others from right to left in the same context (Farghaly & Shaalan 2009). The example below illustrates the Arabic language direction for letters and numbers.

أسست الجامعة البريطانية في دبي في عام 2003

The modules of QA system are developed and adapted to solve language-specific challenges. Due to the complexity of Arabic language among other languages, so the development of Arabic QA systems is a major challenge compared to other languages.

2.5.1. Arabic Script and Encodings

Arabic language uses a particular alphabetic writing system that is considered one of the Arabic challenges. The Arabic common encodings are Windows CP-1256 (One-byte encoding) and Unicode (Two-byte encoding). These encodings are compatible with human's language by allowing them to read and write Arabic texts. While QA systems may encounter many issues when processing Arabic text encoded by one of the codecs mentioned earlier. Due to these issues, some researchers in Arabic NLP prefer to utilize the Buckwalter encoding (ASCII transliteration scheme), which matches Arabic letters to Roman letters as shown in Figure 3. The use of this encoding makes the machine more efficient because of machines designed to work in Roman letters. For example, the Buckwalter transliteration for the text " أســسـت الجامعة البريطانية في دبي في عام 2003" is "Osst AljAmEp AlbryTAnyp fy dby fy EAm 2003" (Benajiba & Rosso 2007).

Arabic letters	ا	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ي	ى[3]
DIN 31635	ʾ / ā	b	t	ṯ	ǧ	ḥ	ḫ	d	ḏ	r	z	s	š	ṣ	ḍ	ṭ	ẓ	ʿ	ġ	f	q	k	l	m	n	h	w / ū	y	ī
Buckwalter	A			v	j	H	x		*				$	S	D	T	Z	E	g								w	y	Y
Qalam	' / aa			th			kh		dh				sh					`	gh									y	
BATR	A / aa			c			K		z'				x					E	g								w / uu	y	ii
IPA (MSA)	ʔ, aː	b	t	θ	dʒ g ʒ	ħ	x	d	ð	r	z	s	ʃ	sˤ	dˤ	tˤ	ðˤ zˤ	ʕ	ɣ	f	q	k	l	m	n	h	w, uː	j, iː	

Figure 3. *Buckwalter transliteration table*

2.5.2. Arabic Morphology

Arabic language is a derivational and inflectional language that makes the analysis of morphology extremely complicated. Some words in Arabic language are composed of one word but are translated into a complete sentence in other languages (Shaalan et al. 2004). For example, the phrase "*Then they will say it*" can be presented in one Arabic word "*فسيقولونها*". Buckwalter (2004) proposed classical morphological analysis (BAMA) which is useful in some simple IR systems but ineffective in the advanced ones. For example, "*من المخترعان اللذان صنعا الطائرة*" and the English translation is "*Who are the two people that invented the airplane*", so the question is looking for the name of two people. In English, the QA system identifies the number of people through the word "Two", while in Arabic the number is embedded in the dual form of the word itself "*شخصان*" (two-persons).

Recently, Pasha et al. (2014) have produced a valuable tool in NLP called MADAMIRA tool, which is a combination of MADA (Morphological Analyzer and Disambiguation tool) and AMIRA (Tokenization, PoS tagging, Phrase Chunking, and NER tool). It designed to process MSA text, as well as the Egyptian dialect as well. In addition, Abdelali et al. (2016) have produced an accurate Arabic text processing called Farasa. It performs efficiently in IR and MT tasks. Farasa consists of several modules, such as Arabic text Diacritizer, PoS tagger, and Dependency Parser.

Regarding the derivations of the Arabic language, each Arabic word has a root of three characters and in some cases four or five characters. The word derivations (Lemma = Root + Pattern) often add a pattern (prefix, infix, or suffix) to the root (Shaalan & Raza 2009). Also, the adjectives and nouns are derived from a verb. In regular derivation, the lemma can be derived if the root and pattern are known. Figure 4 demonstrates a sample of two Arabic verbs from the same classification and their deduction utilizing the same root. The word root or stem plays a key role in the efficiency of IR and QA systems at the stage of indexing or retrieval passages (Harmanani, Keirouz & Raheel 2006).

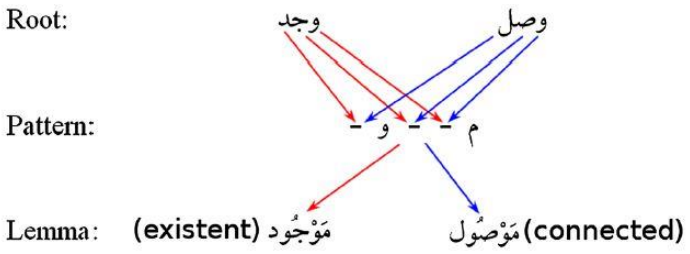

Figure 4. *Arabic language derivation (Lemma = Root + Pattern)*

On the other hand, the Inflectional aspect in the Arabic language is a challenge as well. The word structure consists of the root and add affixes to it (Word = prefix(es) + lemma + suffix(es)). The prefixes can be conjunctions, articles or prepositions, while the suffixes are usually objects or possessive anaphora. Figure 5 illustrates example of prefixes and suffixes that can be combined, so one word can contain zero or more affixes (Harmanani, Keirouz & Raheel 2006).

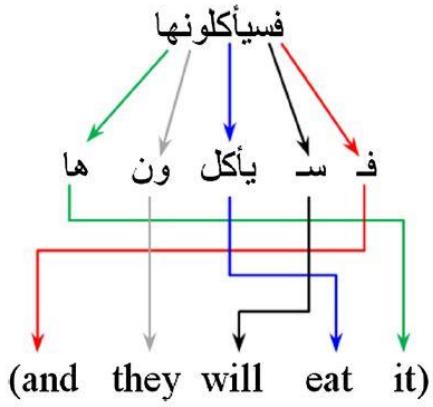

Figure 5. *Arabic words composition (Word = prefix(es) + lemma + suffix(es))*

2.5.3. Natural Language Processing

2.5.3.1. Ambiguity

Unlike Latin languages, diacritics, such as fatha, dama, and kasra signs, play an important role in Arabic language to distinguish either double consonants or short vowels. Figure 6 illustrates the possibility of writing a similar word with different diacritics that gives different meanings (Vergyri & Kirchhoff 2004).

Seriousness	جِد
Grandfather	جَد

Figure 6. *Arabic diacritics changes the meaning*

Regardless of religious text (Quran) and children's books, diacritics are excluded in all Arabic texts to allow Arabic speakers to read faster. In the case of an ambiguous word, ambiguity can be easily identified by utilizing the context in which the word has appeared and reader's knowledge. However, researchers in NLP field consider this point as one of the major challenges, especially in the tasks of Machine Translation (MT) and Word Sense Disambiguation (WSD) (Habash & Rambow 2007). Farghaly & Shaalan (2009) stated that the problem of ambiguity poses a major challenge for researchers to develop Arabic NLP systems. Therefore, non-use of diacritics is an additional challenge. It is estimated that the average number of ambiguities in most languages is 2.3, while 19.2 in Arabic (Cheddadi 2014).

2.5.3.2. Capitalization and NER

Thus, capitalization is one of the main obstacles in Arabic NLP systems because of the difficulty in distinguishing between NEs and the other word which affect NER performance. For instance, the question "أين ولد كريم الذهبي؟" (Where was Kareem Aldahabi born?) asks about the birthplace. A morphological analyzer can interpret the two NE words as an adjective (Kareem Aldahabi interpret into the generous and golden, respectively). Figure 7 presents an example of Arabic capitalization challenge (Shaalan 2014).

Egypt	مصر
teacher	مدرس

Figure 7. *No capital letters in Arabic script*

2.5.3.3. Available Resources for Training

Lack of available or non-free resources is one of the obstacles faced by Arabic NLP researchers. Resources are an essential component for researchers in the research and development cycle. There are many resources used for training purposes such as electronic lexicon, dictionaries and corpora, as well as the advanced ones such as knowledge base and anthology. Therefore, most of researchers build their own resources that are used to train their QA systems. Recently, huge

efforts have been made in development of public resources, particularly those belonging to the electronic lexicon, dictionaries and corpora. Ontology can play a key role in the QA systems, such as Arabic WordNet which has information of Arabic language (Abouenour, Bouzoubaa & Rosso 2013).

2.5.3.4. Available Resources for Evaluation

The development of QA systems goes through many stages and experiments that require relevant resources for evaluation. Currently, in Arabic language there are few test-set available for QA and public domain as well (Shaheen & Ezzeldin 2014). The Arabic language resources available for evaluation as follows: (i) TREC (2001 and 2002) text collections for Arabic QA and IR, which contains 873,383 documents (800 MB). The size of the collection is small compared to the English versions TREC WT10g and TREC GOV2, which contains 1.6 million documents (10 GB) and 25 million documents (420 GB) respectively (Abouenour 2011); (ii) Benajiba, Rosso & Soriano (2007) have created and published an Arabic collection to the public. It was developed in SGML format for their system called ArabiQA system. This collection contains 200 pairs of questions and answers and 11,000 documents from the Arabic Wikipedia; (iii) Trigui, Belguith & Rosso (2010) have produced ADQA corpus (Arabic Definition Question Answering), which was collected from Arabic TV program called “Who will be a millionaire?”. The test-set consists of 50 Arabic definition questions, as well as 50 files collected from Wikipedia and Google for these questions; (iv) Peñas et al. (2011) have provided QA4MRE (Question Answering for Machine Reading Evaluation), which is launched by CLEF 2011. The test-set consist of four topics "Alzheimer", "AIDS", "Music and Society", and “Climate Change”. The content composes of 4 reading tests per topic (16 documents), 10 question per reading test (160 questions), and 5 answers options to each question (800 answers).

2.6. Question Answering Approaches

2.6.1. Linguistic Approach

QA systems need to know how to recognize natural language text to interpret the question and retrieve the answer. Therefore, the development of QA systems relied on AI methods, which utilize NLP techniques and information sources (knowledge base, corpus, etc.) in building system

framework. Information is shaped in the form of rules, templates, ontologies, and semantic networks, which are used in the analysis of questions and answers (Oudah & Shaalan 2016). Linguistic techniques like PoS tagging, parsing, and tokenization were implemented in order to formulate a query from the input question and retrieve the answer from information source. However, there are many limitations in deploying a close-domain knowledge base, because each domain has its own grammars and rules that are used by the system. Whereas, deploying open-domain knowledge base is a time-consuming process and requires a huge effort. Therefore, available QA systems have tried to resolve issues that related to close-domain information (Dwivedi & Singh 2013).

Moreover, the development of some QA systems relied on the rule-based mechanism. The rules were built to be used by NLP techniques in order to identify and classify the features of questions. Quarc and Cqarc systems were developed by Riloff & Thelen (2000) and Hao, Chang & Liu (2007) respectively. Both groups used heuristic rules to identify the class of questions by looking for semantic and lexical clues in the question. However, there is a difference in a class taxonomy of question from one system to another. For instance, some systems use general taxonomy (who, when, what, where and why) while others use domain-specific taxonomy. Table 2 parents NLP-based and Rule-based approaches differences that applied by QA systems.

	NLP-based	Rule-based
Data Structure	Structured Data	Structured or Semi- Structured Data
Question Classification	NLP tasks	Predefined rules
Linguistic Analysis	Deep NLP Techniques	Shallow NLP techniques
Knowledge Domain	Often Small, but large when use web as data source	Limited to pre-stored documents
Learning Data	Not required	Learning rules on training data

Table 2. *NLP-based and Rule-based comparison*

2.6.2. Statistical Approach

The statistical approach has become increasingly important due to the rapid growth of web data and online text repositories. This approach can handle a large amount of heterogeneous data, as well as queries that are formulated in the form of natural language. It relies on statistical learning, which essentially needs a sufficient amount of data to obtain a properly learned statistical program or method. However, this approach fails to identify the linguistic features of the combination words or phrases due to each term is addressed independently and this is one of the

major drawbacks. Indeed, statistical techniques are applied successfully to different modules of the QA system. There are some tasks used to classify the question in order to predict the answer type such as Bayesïan classifiers, Support Vector Machine (SVM) classifiers, and Maximum Entropy models. In addition, these tasks are trained on an annotated corpus (questions or documents) with the classes of questions defined in the system (Dwivedi & Singh 2013).

There are many leading systems developed based on the statistical approach. One of these was IBM's statistical QA system, which used the maximum entropy model to classify questions and answers based on N-gram or bag of words features. The Okapi method was utilized in the IR phase to measure the similarity between question and documents/paragraphs from the data source. While the answer selection phase utilized the heuristic distance metrics to extract the answer (Ittycheriah et al. 2000). Berger et al. (2000) applied the statistical approach in their QA system, as well as the overall performance was quite well and was assessed based on some factors such as vocabulary size, characteristics of the dataset, an overlap between question and answers, etc. Statistical methods such as camel similarity models, N-gram mining, and Okapi are used to determine which documents or answers are closest to a question based on different features and formulas. Moreover, the answer validation module can use the statistical tasks through the appropriate feedback mechanism. Table 3 presents the various statistical models used to classify questions and extract answers, as well as the performance of each model.

Techniques/Models	Phase	Performance
Maximum Entropy	Question Classification	Error rate significantly reduces
SVM	Question Classification	Quite good performance and accuracy
	Answer Selection	Outperforms than others
Bayesïan classifiers	Question Classification	Better accuracy than base Bayesïan method
Okapi Similarity	Answer Selection	Achieved an average level of mean (Factoid questions/Close-Domain)
Sentence Similarity	Answer Selection	Accuracy is significantly increased and precision above average
N-gram Mining	Answer Selection	Satisfactory performance (All type of questions)

Table 3. *Statistical models comparison*

2.6.3. Pattern Matching Approach

Actually, this approach is used by many QA systems that formulate questions into predefined pattern and match them with corresponding answers pattern instead of using linguistic techniques. For instance, the question "*Where was Sheikh Zayed born?*" matches the question's pattern "*Where*

was <Name> born?" and its answer's pattern "*<Name> was born in <Location>*". The simplicity of this approach makes systems very convenient for Factoid questions, which cannot handle the definition questions that require a lot of time and effort. The answer extraction stage used either a surface patterns or templates approach. Table 4 presents a comparison between surface patterns based and templates based (Al-Shawakfa 2016).

	Surface Pattern based	Template based
Mechanism	Automatically learned patterns by examples	A preformatted template for questions that have place-holder to be dynamically filled by parameters
Answer Extraction	Statistical techniques or data mining measures	Structured query
Answer Representation	Not necessarily generating formatted answers	Generating formatted answers
Pattern Learning	Semi-automatic	Manually and automatic for semantic web
Implementation Area	Small/Medium size websites	Semantic web

Table 4. *Pattern matching approach*

2.6.3.1. Surface Pattern Based

The approach gives a high accuracy in the final results, which relies on the list of patterns to extract the answer from the surface structure of documents. It is convenient for Factoid questions since their answers are limited, as well as the answer extracted based on similarities between patterns (e.g. regular expressions) and semantics. This approach has been proposed in the QA evaluation track (TREC- 10), which prompted Ravichandran & Hovy (2001) to adopt this approach in their work. They have applied automatic learning method (bootstrapping) to build a large patterns list of question-answer pairs. Moreover, Zhang & Lee (2002) augmented this approach with confidence measures by giving high precision with a low recall.

2.6.3.2. Template Based

This approach focuses on the illustration using pre-formatted question and answer template rather than processing questions and answers. The template design contains empty slots (placeholder), which are populated based on the question class and generate the query as well to retrieve the answer from the data source. Also, the answer is populated in its template in order to return raw data in a formatted manner. The principle of this approach is similar to the automated Frequently Asked Questions (FAQ) systems, while answers are pre-stored in the templates that

belongs to the user's question. However, this approach unlike static FAQs in the answer retrieval mechanism, where question and answer templates are populated dynamically using parameters. Gunawardena et al. (2010) introduced the QA system for close-domain to identify mobile SMS. They design templates to cover all variant of each potential question. The system relies on pre-processed text to find out the best matched of the template pair (question and answer), and then populate the answer's template accordingly from data source. In addition, Unger et al. (2012) have used this approach over Resource Description Framework (RDF) by using the SPARQL template. The SPARQL template analyzes the question's structure and maps it to domain vocabulary, which gives a fully adaptable to the semantic web. This system also uses linguistic resources to build the SPARQL template.

2.7. Existing Arabic QA Systems (Related Work)

In the 1990s, work began on Arabic IR systems with limitations at that time regarding the resources (size of text collections). Therefore, the systems evaluation focused on measuring the effectiveness of indexing paragraphs by stem, root or surface words. In 2001, the TREC campaign comes with 75 queries to evaluate Arabic IR systems (Nwesri 2008). The first Arabic QA system created was "AQAS", which is knowledge-based system retrieves answers from structured data. The frames mechanism was utilized to present the knowledge in radiation domain. It highlighted some limitations regarding processed data that are structured in the context of knowledge-base. However, there is no available published evaluation of the system (Mohammed, Nasser & Harb 1993).

Ten years later, Hammo, Abu-Salem & Lytinen (2002) proposed the Arabic QA system (QARAB). It is a rule-base system that utilizes a collection of Arabic newspaper texts (unstructured data) to retrieve answers based on IR and NLP techniques. It only processes the factoid questions and does not handle other types of questions like "why" and "how", due to advanced and complex processing needed. However, the evaluation test was carried out by four native Arabic speakers and did not follow the state-of-art methods. It only reported the results of the experiments for 113 questions, which showed that the performance result was obtained for Recall and Precision is 97.3 %.

Benajiba, Rosso & BenedíRuiz (2007) introduced the ArabiQA system, which handles the factoid questions. The NER and Java Information Retrieval System (JIRS) techniques were used to process the Arabic text documents and factoid questions. The authors prepared a system of ongoing implementation and an evaluation corpus based on the CLEF guidelines. The obtained accuracy of system performance is 83.3%.

Brini et al. (2009) built the Question Answering System for Arabic Language (QASAL), which handles the factoid questions that have NE answers. The NooJ platform and local grammars were used to retrieve the answers. As well as the IR and NLP techniques were used for processing Arabic text documents and factoid questions. The authors utilized the Arabic version of Google as a data resource, as well as used the Tunisian book collection as a corpus. According to the authors, the results of experiments that conducted on 50 questions showed the 67.65% Precision, Recall is 91% and F-measure is 72.85%.

Kanaan et al. (2009) introduced the Arabic QA system to handle the short factoid questions, while other types of questions such as "why" and "how" are excluded. IR and NLP techniques were used to retrieve answers from variety of data source, such as Arabic text collection, some relevant documents, and 25 documents manually collected from the Internet. According to the authors, 12 questions were applied to evaluate the system. The reported performance shows the Recall is100% and the Precision obtained is 43%. Due to the small size of the sample used, the results of their experiments cannot be relied on in comparative.

Trigui, Belguith & Rosso (2010) were introduced the Arabic Definition Question Answering system (DefArabicQA), which handles the question's pattern “What is X?” and use the internet as a data source. Google and Wikipedia were used in the experiments as a data source to extract answers. The reported results: MRR score is 0.7 and the question rate is 0.54 for Google as the web source, while MRR score is 0.81 and the question rate is 0.64 for Google coupled with Wikipedia as the web source.

Bekhti et al. (2011) built the AQuASys system, which handles factoid questions in order to retrieve answers from Arabic corpus that developed by the authors themselves. It consists of three modules: Question analysis, Sentence filtering and Answer extraction as well as extensive use of

NLP techniques. According to the authors, the results obtained from the experiments on 80 questions showed a 66.25% Precision, Recall is 97.5% and F-measure is 78.89%.

Trigui et al. (2012) have introduced the Arabic QA4MRE system, which introduced the Arabic language for the first time in CLEF. The system adopts a new approach that can answer questions of multiple answer options from short texts. The reported performance shows that the overall calculated accuracy is 0.19 (31 correct questions answered correctly out of 60) and the overall c@1 measure is 0.19 as well.

Abdelnasser et al. (2014) have established Al-Bayan system, which is a QA system for the Holy Qur'an that uses Qur'an and its interpretation books to answer questions. The structure of the system consists of three modules, a Semantic module to restore the verses from Qur'an, Morphological analyzer and disambiguation to classify questions, Answer extraction to rank answers from interpretation books. According to the authors, the system accuracy is about 85%.

Kurdi, Alkhaider & Alfaifi (2014) have proposed new solution called JAWEB, which is a web-based Arabic QA system that handles factoid questions (Yes/No questions) based on paragraph retrieval. The experiments were conducted on small size collection (20 documents). The system shows positive results of about 85% Precision and 100% Recall. Besides, the system gives 88% Precision when using only paragraphs.

Nicosia et al. (2015) have introduced Answer Selection in Community Question Answering for Arabic and English languages. The tool targeted web forums, where posted questions and their comments were used to select the best answer. ML approach was adopted to classify the question's comments into true and false answers. Moreover, a variety of features, such as sentiment analysis, n-grams, and text similarity. According to the authors, the tool accuracy is about 78.69%.

AL-Khawaldeh (2015) has presented an Arabic QA system (EWAQ) based on entailment metrics that handle Why-questions. It is designed to use the web as a data source and score the answers based on entailment metrics to find the best correct answer. The system shows positive results of about 68.53% Accuracy for 205 Why-questions.

Albarghothi, Khater & Shaalan (2017) have introduced Arabic QA system based on ontology (Domain knowledge). It is designed based on NLP tasks, Protégé tool, and SPARQL queries. The experiments were conducted on 100 questions (76 answered questions, 18 incorrect answer, and 6

questions blank answer). The system shows positive results of about 81% Precision, 93% Recall, and 86% F-Measure.

Azmi & Alshenaifi (2017) have introduced an Arabic QA system (Lemaza) that handle Arabic Why-questions. The system employed the Rhetorical Structure Theory (RST) to extract answers. The experiments were conducted on 110 Why-questions using 700 documents compiled from open source Arabic corpora. The system shows positive results of about 72.7% Recall, 79.2% Precision, and 78.7% c@1.

Most of Arabic QA systems that handle type of question (Factoid or definitional) seek for a precise and straightforward answer. Only a few of the researchers introduced a project managing definitional questions (Why and How) due to their difficulties and complexities (Al-Shawakfa 2016).

2.8. Existing Arabic Corpora (Related Work)

In recent years, computational linguistics and natural language have evolved considerably, including changes in QA systems such as the types diversity and complexity of questions, as well as standardized evaluations such as TREC QA track. There is a major shift from developing simple systems to comprehensive systems by using large corpora. Some approaches of QA systems have benefited from external resources such as gazetteers, web resources, encyclopedia, and databases (Kennedy 2014).

Since NLP studies cannot rely on small samples of data as well as the intuition, so they require a vast data (corpus-based approach) to perform the experimental analysis. The corpus-based approach can be utilized to conduct a study on a wide range of topics in NLP field. Corpora are extremely appropriate for QA functionally because it comprises of texts that empower the researchers to develop systems for their tasks analyses (Alansary, Nagi & Adly 2007).

2.9. Chapter Summary

In this chapter, we have highlighted the importance QA systems to the users and the aim that attract the attention of researchers, as well as we have identified the difference between these systems and SEs. Typically, QA systems consist of three major modules: Question analysis, Passage retrieval, and Answer extraction. While each module is integrated with a variety of NLP

tasks. Generally, the performance of these systems is evaluated using a range of tests, such as the mean of IR systems measure by recall and precision formulas, as well as QA-specific measure by accuracy, MRR and C@1 formulas.

Although the architecture design of QA systems and the evaluation process are independent of the language. While the core tasks in Arabic language have been developed to solve some of the challenges that belong to this language. These challenges are illustrated in this chapter along with examples for further clarification. Early studies on Arabic QA systems do not report the experimental results based on standard measures, since Arabic was not introduced in current campaigns (provided only in TREC 2002 and CLEF 2012 / 2013).

Moreover, we reviewed in detail the most prominent Arabic systems through the system approaches, functions, and results in order to identify the new lines of our research. Therefore, we found that the development of Arabic QA systems is highly concerned due to the availability of resources as well as tasks such as QE, NER, syntactic parsing, etc. Finally, we reviewed the Arabic corpora in terms of approaches and the most prominent works.

Chapter Three: Building Corpus

This chapter explains the reasons beyond the creation of our linguistic resource Arabic Corpus of Occupations (ACO) corpus. The linguistic resource has a great impact on our research, so it highlights the methodology of building ACO corpus that will be required to train and test our QA system. Finally, it demonstrates the evaluation process and its factors that will be used to assess our corpus.

3.1. Motivation

As mentioned earlier, Arabic language still lacks the large-scale of resources, and this is necessary to implement NLP solution. Therefore, researchers have become more interested in this subject by achieving significant progress. Arabic NLP is still in its earliest stages for many reasons, one of them is the lack of a mature data resource. Hamoud & Atwell (2017) pointed out that the development of the QA system for close-domain has a wider interest than open-domain, due to the integration of information that related to the selected domain. Thus, we choose the occupation domain to build the corpus and utilize it as a data source for our QA system for the following reasons: (i) covering the diverse range of jobs and their descriptions in one place based on the International Standard Classification of Occupations (ISCO-88) for the classification and compilation of occupational information; (ii) provide a language of common understanding regarding the professional structures of the labor force; (iii) no existing resources specifically designed for the occupations domain; (iv) NLP studies require a vast data (corpus-based approach) to perform the experimental analysis. Thus, ACO is a valuable resource for QA development (close-domain) by creating a unified dataset for use in testing and evaluation purposes.

3.2. Corpus Aims

In order to make the corpus more useful for QA system, it is often subject to annotation procedure, such as semantic analysis, morphological analysis, and statistical analysis (Hamoud & Atwell 2017). The ACO is planned to have a sizable and reliable Arabic content written in MSA format collected from appropriate sources. ACO aims to provide a well-structured guide for building and processing Arabic corpora, as well as produce an annotated Arabic corpus using a reasonable tag set. Producing an ACO corpus requires data collection, human annotation, and

verification. Therefore, a customized tool will be developed to collect and process Arabic corpus in order to add positive progress in this field.

Moreover, ACO is a valuable data source for our QA system, which will be used in testing and evaluation experiments. Moreover, our QA system (employing passage retrieval module) can be utilized as a recommender system by matching users' resumes with the most appropriate occupations.

3.3. Methodology

The designed methodology helps us in creating a reliable corpus. The approach of manual annotation (tagging) will be chosen because it is more accurate but requires a considerable time and huge efforts. Figure 8 illustrates the ACO methodology cycle which consists of (i) define the knowledge domain and choose the most appropriate resources, as well as identify the most suitable data sources; (ii) design the structure of the corpus, written properly and carefully organized (headings, sections, paragraphs, etc.); (iii) clean, upload, reformat data by eliminating noise and adding tags/metadata (annotations) to improve the data accuracy; and (iv) perform a baseline assessment in order to calculate the accuracy.

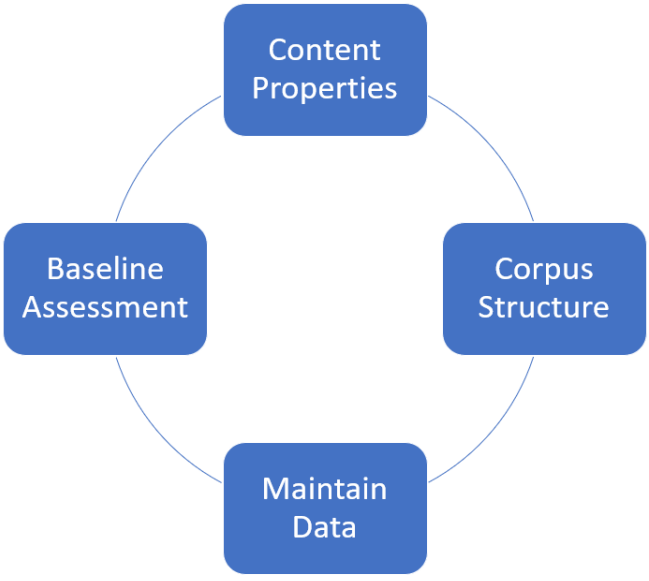

Figure 8. *ACO development methodology*

3.4. Planning the Building of Corpus

In the planning phase, various factors should be considered for building the corpus, such as sampling and representativeness, machine-readable form, finite size, and status as a standard reference (Goweder & De Roeck 2001). Therefore, the sampling and representativeness factor can be relied upon to formulate research questions of research. List of the following tasks have been identified to be followed in the development and compliance process of the corpus: (i) identify the data/text domain; (ii) structure or classify the text representation; (iii) identify resources and any roadblocks; (iv) maintain a precise record such as text source, tag, last update, user, etc.; (v) identify the methods of training and testing evaluation; (vi) Identify the categories, the samples in each category, and the max number of words in each sample (for corpus size purposes).

3.5. Data Collection

ACO is planned to have approximately one million words written in MSA format. The collected data consist of a diverse range of occupations as part of a funded project with grant number 37-K-138 supported by Princess Noura Bint Abdulrahman University, Riyadh, Kingdom of Saudi Arabia. The project called Electronic Dictionary of Professions (المعجم الحاسوبي لمصطلحات المـهـن), which aims to represent all job descriptions and their respective groups in the Arabic language. The data structure is made up of major groups (الاقسام) at the top level, subdivided into sub-major groups (الاجزاء), subdivided into minor groups (الابواب), and unit groups (الفصول).

3.6. Corpus Analysis and Design

The proper design of the corpus relies on what is intended to represent. ACO has adopted the International Standard Classification of Occupations (ISCO-88) for the classification and compilation of occupational information. The standard defines and aggregates jobs based on the similarity of the skills required to fulfil the duties of these jobs. The aim of this standard is to provide a language of common understanding regarding the professional structures of the labor force among all countries of the world.

It is important to note that building the ACO is a hazardous errand because of covering the diverse range of jobs and their descriptions. The ACO framework adopted two basic concepts of classification, namely: type of work performed (Job) and skill. The job is defined as the set of tasks

and duties performed by a person. While the work is part of the occupation, which includes a set of tasks homogeneous in nature but different in skill level. On the other hand, the skill is defined as the ability to perform tasks and duties precisely according to the requirements of the labor market. The skill has two dimensions: (i) skill level which is the scope and complexity of the tasks involved; (ii) skill specialization which includes the type of knowledge applied, tools and equipment used ("ISCO - International Standard Classification of Occupations" 2018).

Error! Reference source not found. shows the five levels of skills were delineated in ACO c orpus. Whereas, **Error! Reference source not found.** shows the ACO's hierarchical of 10 major groups at the top level, subdivided into 40 sub-major groups, subdivided into 142 minor groups, and 430 unit groups.

فئات مستوى المهارة	الرمز	الوصف
فئة مستوي الاختصاصى	1	تشمل فئة مستوي الاختصاصي الأعمال التي يتطلب إنجازها توافر قدر عال من المهارات المعرفية والتقنية والإدارية لدى شاغليها لتمكينهم من تحسين وتطوير المبادئ والمفاهيم والطرائق والأساليب الإجرائية، وتطبيق حصيلة المعرفة العلمية والمعرفية في مجال الشغل، ولتمكينهم أيضا من متابعة العاملين في أثناء التنفيذ، وتقييم الإنجاز، وحل مشكلات العمل والعاملين. ويحتاج الأفراد الذين يشغلون أعمالا ضمن هذه الفئة إلى إعداد وتأهيل جامعي ومن الأعمال التي تصنف في هذه الفئة: محام، ومهندس مدني/ أبنية، وطبيب، واختصاصي اجتماعي، وطبيب اختصاصي
فئة مستوى الفني (التقني)	2	تشمل فئة مستوى الفني التقني الأعمال التي يتطلب إنجازها تطبيق المبادئ والمفاهيم والطرائق والأساليب الإجرائية ذات الصلة بالشغل. ويتطلب هذا توافر مهارات علمية وفنية وأدائية وإشرافية لدى شاغلي الأعمال ضمن هذه الفئة، لتمكينهم من فهم طبيعة الأداء وتحليله، وتحديد خطوات الإنجاز ومتابعة تنفيذها وتقييمها، ويمثل العاملون في هذه الفئة حلقة الوصل بين الاختصاصيين والعاملين. يحتاج الأفراد الذين يشغلون أعمالا ضمن هذه الفئة إلى إعداد وتأهيل متوسطين (ما بعد التعليم الثانوي ودون التعليم الجامعي)، أو في مستوى كليات المجتمع أو ما يوازيها. ومن الأعمال التي تصنف في هذه الفئة: فني مختبر مواد، ورسام معماري، وفني كهرباء
فئة مستوى العامل المهني	3	تشمل فئة المستوى المهني الأعمال التي يتطلب إنجازها توافر مهارات عملية ومعلومات مهنية تغطي إطار المهنة بشكل متكامل لدى شاغليها، لتمكينهم من ممارسة مهام وواجبات المهنة وبدرجة إتقان بحسب متطلبات سوق العمل. ولتمكينهم من توزيع العمل على المرؤوسين وتنمية مهاراتهم. يحتاج الأفراد الذين يشغلون أعمالا ضمن هذه الفئة إلى تعليم أو تأهيل مهني بعد إنهاء مرحلة التعليم الثانوي كأساس لمدة قد تصل إلى عام تدريبي. ومن الأعمال التي تصنف في هذه الفئة: خراط عام، وممرض عملي، وطابع عام، وكهربائي تمديدات كهربائية
فئة مستوى العامل الماهر	4	تشمل فئة مستوى العامل الماهر الأعمال التي يتطلب إنجازها توافر مهارات عملية ومعلومات مهنية تتصل بجزء من المهنة، ولا تغطي إطار المهنة كلها لدى شاغليها، لتمكينهم من أداء مهام العمل وواجباته وبدرجة إتقان بحسب متطلبات سوق العمل. يحتاج الأفراد الذين يشغلون أعمالا ضمن هذه الفئة إلى تعليم أو تأهيل مهني يوازي مرحلة التعليم الثانوي. ومن الأعمال التي تصنف في هذه الفئة: لحيم كهرباء، وكهربائي تمديدات منزلية، وطابع/ لغة إنجليزية، ومربي دواجن

فئة مستوى العامل محدود (محدد المهارات)	5	تشمل فئة محدود (محتد) المهارات الأعمال التي يتطلب إنجازها توافر مهارات عملية ومعلومات مهنية تغطي جزءاً ضيقا من المهنة، أو الأعمال التي تشمل مهاما وواجبات روتينية يتطلب إنجازها استخدام أدوات يدوية محدودة ومجهود عضلي، ولا تحتاج إلى معلومات مهنية تتعلق بالأدوات والمواد والمنتجات. ويمكن اكتساب هذه المهارات إما عن طريق التدريب القصير أقل من تسعة شهور) أو من الخبرة، أو التعلم الذاتي، أو التدريب في موقع العمل. ومن الأعمال التي تصنف في هذه الفئة: مشغل آلة، وبائع صحف، ومساعد كهربائي تمديدات منزلية، ومساعد لحيم كهرباء، ومصلح إطارات مركبات، وبواب عمارة، وحارس

Table 5. *Skill levels*

الاقسام (المجموعات الكبرى)	الاجزاء (المجموعات الفرعية)	الابواب (المجموعات الصغرى)	الفصول (المجموعات القاعدية)	مستوى المهارة
المشرعون وكبار الموظفين والمديرون	5	10	40	*
الاختصاصيون	6	25	73	1
الفنيون ومساعدو الاختصاصيين	6	22	71	2
الكتبة	4	8	21	3,4,5
العاملون في مهن الخدمات والبيع في الأسواق والمحلات	4	11	34	3,4,5
العاملون في الزراعة وصيد الأسماك	2	5	15	3,4,5
الحرفيون والمهن المرتبطة بهم	4	18	80	3,4,5
مشغلو المصانع والآلات وعمال التجميع	3	19	66	3,4,5
العاملون في المهن الأولية	3	11	27	5
العاملون في القوات المسلحة	3	3	3	*

Table 6. *ACO's hierarchical structure*

ACO uses the decimal coding system for the job reference as shown in Figure 9, which consists of seven digits representing six classification boundaries as follows:

- The first slot allocates one digit (the first digit from the left) and represents the main groups.
- The second slot allocates one digit (the second digit from the left) and represents the sub-major groups.
- The third slot allocates one digit (the third digit from the left) and represents the minor groups.
- The fourth slot allocates one digit (the fourth digit from the left) and represents the unit groups.

- The fifth slot allocates two digits (the fifth and sixth digit from the left) and represents the jobs.
- The sixth slot allocates one digit (the seventh digit from the left) and represents the skill levels.

Job Code	Job Name
2111011	فيزيائي / عام

Figure 9. *Job reference sample*

3.7. Arabic Corpus of Occupations Tool

According to the previous analysis and design section, the ACO tool is an appropriate application to help the user explore and tag (annotate) the Arabic occupations. The tool has been developed using C# windows application and Microsoft Office 2016 (Access and Excel). It consists of two different parts, one to fill the data in the corpus and the other to verify and tag (annotate) this data. The tool has a variety of features and capabilities such as data insertion and a workflow to annotate and verify content, as well as ease of use and allows the user to work smoothly. Figure 10 shows the interface of ACO tool.

Figure 10. *ACO tool*

3.8. Insertion and PoS Tagging Process

This corpus has the advantage of being manually tagged which ensures its cleanliness and accuracy. The manual approach is followed to tag around 700 occupations of the collected data. The process of insertion and tagging consists of three different stages, namely: (i) the tagging (annotation) stage and carried out by tagger role who fills and annotates the information; (ii) the verification stage and carried out by domain specialist role; (iii) the approval stage and carried out by expert role. As illustrated in Figure 11, the tagging stage begins when the tagger role fills in the required information as well as adding the corresponding tag from the tag set. In the verification stage, the reviewer role (domain specialist) verifies the inserted content as well as its annotations. The reviewer may update the inserted content and mark the row as "Verified with update" state. In contrast, if there is no content update, the row will be marked as "Verified" state. In the approval stage, the expert role approves the inserted content. In some cases, the expert may ask for amendments and return the content to domain specialist to do the needful.

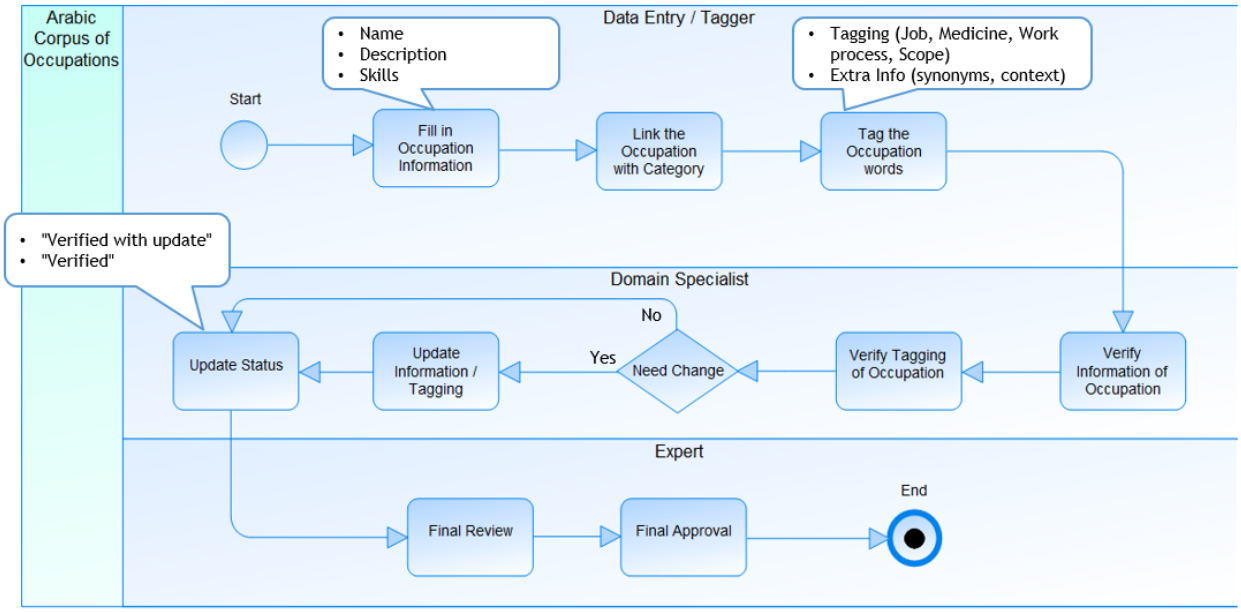

Figure 11. *ACO tagging process*

The set of tagging words were assigned based on Arabic grammar books and verified by an Arabic linguistic specialist. In this paper, Arabic classifications (Job name, Job process, or Job scope) were used in the tagging process. As well as, the synonyms are added to the name of the

job itself, as well as the context in which the name of the job or the nature of the work may come. Table 7 shows the example of job and its associated tags.

Job Name	Synonyms	Context	Tag set
طبيب عام	طبيب عام ممارس	وقد يطلب هذه الطبيب العام تحاليل الدم	الطبيب العام الممارس (اسم وظيفة)
	طبيب في الطب العام	المتلازمات الغدية غير مألوفة للطبيب العام	الالتهاب الفيروسي والبكتيري (اسم مرض)
	طبيب ممارس عام	دائما ما يستشر المريض الممارس العام من اجل علاج بعض التغيرات الفسيولوجية	تشخيص الحالات المرضية العامة للأشخاص من كافة الفئات العمرية ومعالجة هذه الحالات، وتقديم المعالجة الطبية الطارئة للحالات التي تستدعي ذلك، وتحويل الحالات المرضية التي تستدعي تداخلات واستشارات طبية أخرى إلى المختصين، ومتابعة الحالات المرضية المعالجة (الية العمل)
	ممارس عام		دراسة الحالات المرضية العامة والسيرة المرضية للمريض، وإجراء الفحوصات السريرية والمخبرية والشعاعية الأولية وتشخيص الحالات. وضع خطة المعالجة لها ومتابعتها، ومناقشة خطة المعالجة مع المريض أو الأهل، وتحديد العوامل المساعدة في نجاح خطة المعالجة من أدوية وأغذية وبيئة، وتحويل الحالات التي بحاجة إلى الاستشارات والتداخلات الأخرى إلى المختصين (المجال)

Table 7. *Example of job annotation*

3.9. Preliminary Processing

In order to have an error-free corpus, we investigated the content of ACO corpus to verify and cleans a number of issues related to Arabic spelling and orthography. Lists of unique words were generated by executing a customized script that helps us to assess the data. We noticed some noisy data that required a fix. For example, the preposition “on” that misspelled as (علي) instead of (على) and the letter of joining words "wow" that misspelled as one-word (وفي) instead of (و في).

3.10. Corpus Assessment

As mentioned in the planning section, the building of the modern corpus was relied on various factors such as sampling and representativeness, machine-readable form, finite size, and status as a standard reference. Most of the factors were achieved by the ACO corpus except the last point that is not yet a standard reference. A variety of arguments make ACO corpus representative of modern Arabic language. For example, it covers a diverse range of occupations and written by many authors.

We have begun evaluating the data collected after the pre-processing stage in order to obtain high-quality data in a machine-readable format. After that, the assessment process is started by calculating the corpus' statistics in order to figure out the corpus properties in terms of time, amount of data, etc. as detailed in Table 8. Finally, we measured the density of words and sentences that give a detailed view of the data. Table 9 gives an indication of n-gram distribution in the corpus.

Basic Word Count Statistics	
Words	131524
Characters (including spaces)	977938
Characters (without spaces)	846307
Extra Word Count Statistics	
Syllables	131686
Sentences	4451
Unique Words	13076
Average Word Length (char)	6.4
Average Sentence Length (word)	29.5
Monosyllabic Words (1 syllable)	131420
Polysyllabic Words (≥3 syllables)	40
Syllables per word	1
Paragraphs	7
Length Statistics	
Short Words (<=3 characters)	11325
Long Words (>=7 characters)	61345
Reading Time	
Estimated Reading Time	658 min.
Estimated Reading Time	1053 min.

Table 8. *Statistics of ACO content*

Top Keyword Density (1 Word)	
1. في	2570 *(2%)*
2. وتطوير	2239 *(1.7%)*
3. وتحديد	1917 *(1.5%)*
4. والصحة	1801 *(1.4%)*
5. وإعداد	1796 *(1.4%)*
6. السلامة	1703 *(1.3%)*
7. إجراءات	1568 *(1.2%)*
8. إعداد	1378 *(1%)*
9. مستلزمات	1215 *(0.9%)*
10. التقارير	1128 *(0.9%)*
Top Keyword Density (2 Word)	
1. السلامة والصحة	1546 *(1.2%)*
2. وتطوير إجراءات	1240 *(0.9%)*
3. مستلزمات السلامة	1119 *(0.9%)*
4. وتأمين مستلزمات	1116 *(0.8%)*
5. إجراءات وتأمين	1109 *(0.8%)*
6. التقارير الفنية،	916 *(0.7%)*
7. في مجال	753 *(0.6%)*
8. الصلاحية المخولة	631 *(0.5%)*
9. وضمن الصلاحية	629 *(0.5%)*
10. المخولة له	623 *(0.5%)*
Top Keyword Density (2 Word)	
1. وتأمين مستلزمات السلامة	1110 *(0.8%)*
2. إجراءات وتأمين مستلزمات	1108 *(0.8%)*
3. وتطوير إجراءات وتأمين	1084 *(0.8%)*
4. مستلزمات السلامة والصحة	1067 *(0.8%)*
5. وضمن الصلاحية المخولة	629 *(0.5%)*
6. الصلاحية المخولة له	620 *(0.5%)*
7. المخولة له ويقوم	619 *(0.5%)*
8. منفردا وضمن الصلاحية	606 *(0.5%)*
9. وإعداد التقارير الفنية،	543 *(0.4%)*
10. وتوفير الظروف الصحية	540 *(0.4%)*

Table 9. *N-gram distributions*

3.11. Results and Discussion

As stated in the first section of this chapter, the motivation for this work is the lack of free resources as Arabic research is becoming increasingly important. Indeed, very few tagged corpora are available for free, which induced us to build this corpus serving this paper and make it available free for the public. The ACO corpus has been compiled from various sources and tagged manually using set of tags, as well as jobs synonyms and their context are added, which considered the basic step in the linguistic analysis. Later, this corpus is used as a base to build our QA system in order to conduct testing and evaluation experiments. Moreover, our QA system (employing passage

retrieval module) can be utilized as a recommender system by matching users' resumes with the most appropriate occupations.

As pointed out earlier in the ACO tagging process, Arabic grammar books and Arabic linguistic specialist were utilized to verify the tagged information and obtain reliable data. The Cohen's Kappa coefficient method was used to measure the reliability of the tagged information, which computes the degree of agreement level between two taggers (annotators) (Takala et al. 2014).

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Where Pr(a) is the relative agreement among taggers (identical) and Pr(e) is the probability of random chance agreement (each observer randomly).

Table 10 presents the results of data analysis for 2,865 words tagged and assessed by two different groups of taggers. The table highlights the number of agreement and disagreement between them.

		Group (B)	
		TRUE	FALSE
Group (A)	TRUE	1806	39
	FALSE	53	917

Table 10. *Agreement & disagreement matrix*

Table 11 shows the calculation procedure to obtain the K value of the tagged content in the ACO corpus.

Variable	Formula	Result	Percentage
Pr(a)	(TT + FF) / (TT + FF + TF + FT)	0.9673	97%
P(True)	[(TT + TF) / (TT + FF + TF + FT)] * [(TT + FT) / (TT + FF + TF + FT)]	0.4361	44%
P(False)	[(FF + FT) / (TT + FF + TF + FT)] * [(FF + TF) / (TT + FF + TF + FT)]	0.2242	22%
Pr(e)	P(True) + P(False)	0.6603	66%
K	(Pr(a) – Pr(e)) / (1 – Pr(e))	0.9037	90%

Table 11. *ACO Kappa formula results*

Based on the above figures for the Kappa formula results, the obtained outcome was 90%. According to Table 12, the reliability of the ACO corpus is almost perfect agreement (Takala et al. 2014). Therefore, the content of the corpus is highly confidence and reliable according to the result achieved and the interpretation of the table below.

Value of Kappa	Level of Agreement	% of Reliable Data
0 – 0.20	None	0 – 4%
0.21 – 0.39	Minimal	4 – 15%
0.40 – 0.59	Weak	15 – 35%
0.60 – 0.79	Moderate	35 – 63%
0.80 – 0.90	Strong	64 – 81%
Above 0.90	Almost Perfect	82 – 100%

Table 12. *Interpretation of Kappa value*

3.12. Chapter Summary

In this chapter, we have highlighted the importance of the linguistic resource to the IR systems and the aim that attracted us to build the ACO corpus. Besides, we have detailed the methodology cycle which compose of 4 stages, namely: (i) define the content properties and sources; (ii) design the corpus structure; (iii) perform data manipulation; (iv) perform a baseline assessment.

ACO corpus has approximately a million words written in MSA format. The collected data represent a diverse range of occupations as part of a funded project with grant number 37-K-138 supported by Princess Noura Bint Abdulrahman University, Riyadh, Kingdom of Saudi Arabia. We present a corpus of 700 occupations which are analyzed carefully and manually annotated. Moreover, we have produced an appropriate application to help us explore and tag (annotate) the corpus content.

It is important to note that the creation of ACO is a cyclic process that requires continuous evaluation during content compilation. We have used Cohen's Kappa coefficient method to evaluate the reliability of the tagged content. The corpus content has been tagged and assessed by two different groups of taggers. Moreover, the synonyms are added to the name of the job itself, as well as the context in which the name of the job or the nature of the work may come.

Accordingly, the inter-annotator agreement indicates that the reliability of ACO corpus is almost perfect agreement. As well as, the content of the corpus is highly confidence and reliable according to the result achieved by 90%.

Chapter Four: Arabic Question Answering System

This chapter explains the reasons beyond the creation of our QA system. Also, it highlights the methodology of building our system, as well as the process of training and testing the system. Moreover, it illustrates the key components of our system and its tasks. Finally, it demonstrates the evaluation process and the test formulas that will be used to measure the effectiveness and performance of our system using different data set.

4.1. Motivation

As stated in Chapter 2, the most experiments of Arabic QA are still restricted to a simple question (factoid questions). As well as, the development of these systems is still in the early stages to handle a complex type of questions. Therefore, we require new methodologies or approaches that utilize the NLP tools and resources available to develop a new system that may add positive progress in this field. Recently, the efforts exerted in the Arabic NLP community have led to develop many tools and resource. Therefore, this research is an opportunity to adopt and develop a more advanced system.

QA systems are more sophisticated and complex than SEs as they look for a precise answer to the query. The need for Arabic QA systems has become increasingly important because of the growing volume of Arabic content on the web, as well as the demand for precise information as well. The regular IR techniques cannot fulfill this need, which allows the user to retrieve only documents and paragraphs that match a particular query. Therefore, the ability to obtain a concise and accurate answer draws attention to QA systems.

4.2. System Aims

The main goal of our QA system is to handle and answer all types of questions beyond the factoid questions. Achieving this goal will increase the popularity of QA systems among users, mainly the users have used the web content and social media content. Henceforth, we will design a new approach to develop an effective Arabic QA system with the ability to address some of the following challenges: (i) improve the main modules of QA system, such as retrieving and ranking passages considering the Arabic language characteristics; (ii) handle factoid and complex

questions (How and Why) in Arabic language; (ii) extract the precise answer from available resources; (iii) system evaluation based on gold standard data set.

According to previous research conducted in this field, we found that the most important module in the QA architecture pipeline is the passage retrieval. Therefore, the focus on this component will significantly improve the performance of QA systems, as well as guarantee the provision of high-quality passages that belong to the question.

4.3. Methodology

The methodology of this research is based on the experiential-oriented approach which consist of different stages. Figure 12 illustrates our QA system methodology cycle which consists of (i) data preparation in order to have more meaningful and robust data; (ii) design the system algorithms that handle all types of Arabic questions by integrating NLP tools with available resources; (iii) develop the system and conduct the necessary experiments to monitor system performance; (iv) evaluate the effectiveness of the system, especially complex questions such as definition and why questions.

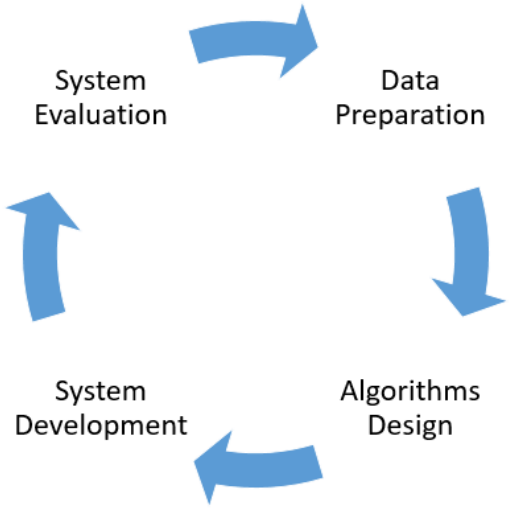

Figure 12. *Our proposed Arabic QA system methodology*

4.4. Data Perpetration

As part of preliminary processing, we performed some normalization tasks on ACO corpus such as strip all the diacritics and replace some letters with others such as replace آ / أ / إ / with ا , ي with ى , and replace ة with ه.

Moreover, QA evaluation campaigns such as CLEF and TREC tracks provide a set of questions with their answers as well as a list of documents to extract the right answer. In order to cover the research's question related to QA system evaluation, we utilize the questions available from different versions of TREC and CLEF to assess the performance of our system. Using these two data sets (Gold Standard) in our experiments helps us to compare our system performance with the baseline system performance. These sets are produced in a variety of languages except Arabic. Therefore, there is a need to translate the content of these sets into Arabic language, so we have manually translated the TREC and CLEF questions and their answers are listed in English to Arabic language.

The total number of translated questions was 800 for CLEF set and 1,500 for TREC set, which were classified in different question's types and domains such as history, politics, sports, etc. Table 13 and Table 14 present the types of questions and the number of questions in each type. As illustrated in both sets, the majority of questions belong to the type of Factoid, which the expected answer is NE. In contrast, the percentage of complex questions is much lower than the Factoid, which are more important for QA systems evaluation.

Q Type	Count
Factoid	498
Complex	187
Other	115
Total	**800**

Table 13. *CLEF Set*

Q Type	Count
Factoid	830
Complex	330
Other	340
Total	**1500**

Table 14. *TREC Set*

4.5. System Architecture

Actually, there is a lot of Arabic QA system outlined in the literature review (Chapter 2). These systems are built based on different techniques and architectures, and it is difficult to adapt all variations in a single architecture. Since most of QA systems have a variety of common features,

it enabled us to use these features in our system design as well as utilize the advantages of NLP tasks. This thesis is part of ongoing research in the field of Arabic QA. Figure 13 illustrates the main components of our QA system and the manners in which they interact. The pipeline structure composed of three major components, namely: Question Analysis, Passage Retrieval and Answer Extraction. While each component consists of several modules with a distinct function.

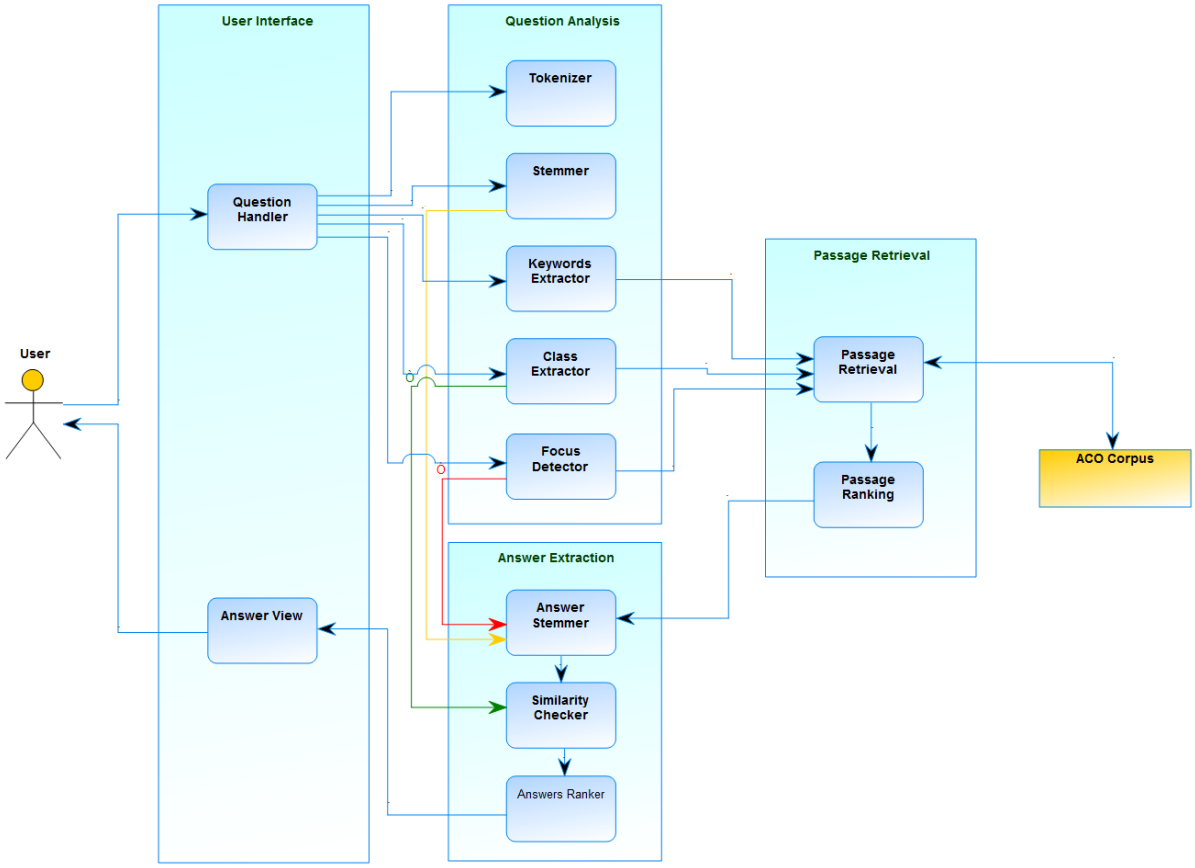

Figure 13. *Generic system architecture*

The aim of our system is to analyze a wide range of question's types, and accordingly identify relevant passages and extract the correct answer from them. The proposed system is a web-based Arabic QA system built from scratch for occupation domain. The question is handled in natural language and extracts the relevant answer from ACO corpus.

Question analysis is the initial stage of any QA system, the component's performance significantly effects on the modules that follow (passage retrieval and answer extraction). In order to get better performance, all semantic features available in the questions should be extracted. Therefore, this component receives the question in natural language as input and performs the morphology (the study of word forms) and syntax (the study of sentence structure) analysis in order to determine syntactical and semantical features. Moreover, it uses some linguistic tasks such as tokenizer, stop-word removal, and stemmer which help to predict and determine the answer type. Accordingly, a passage retrieval query is formulated and forwarded to the next component.

Typically, passage retrieval is a standard component that identifies the relevant passages which are most likely to contain an answer based on the terms of received query. This component returns a list of ranked passages which are forwarded to the answer extraction component for further analysis.

Finally, answer extraction receives the answer type and the list of ranked passages in order to extract the answer using a set of features. This component extracts the candidate's answers and chooses the most probably phrase to be the correct answers.

4.5.1. Question Analysis

This component receives a user question in natural language as input for careful analysis to understand the purpose of the question before sending its information to the IR component. The user's question should be normalized to transform text to harmonic format by performing such steps as removing punctuation marks, removing non-letters, removing diacritics, and removing white-space. The components consist of five modules, namely: (i) Tokenizer, which segment the question into separate words called token; (ii) Stemmer, which removes the suffixes to compare the stemmed keywords with answer later; (iii) Keywords Extractor, which splits the question into an important words after excluding interrogative nouns, conjunctions, prepositions, and stop-words; (iv) Class Extractor, which identifies the question’s class and accordingly the type of expected answer which can be NEs as shown in Table 15; (v) Extra Keywords, which specifies exactly what the question is looking for (such as NEs) and the synonyms of keywords. Figure 14 presents the high-level design algorithm for this component.

Interrogative Particles		Question Type	Scope
متى	When	Factoid	Time
اين	Where	Factoid	Location
من	Who	Definitional	NE (Person)
ما	What	Definitional	Object
مما	What	Definitional	Object
لمن	Whose	Factoid	NE (Animate)
لمن	Whose	Factoid	NE (Person)
كم	How much / many	Factoid	Quantity
كيف	How	Method	Process to do something
لماذا	Why	Purpose	Reason for doing something
اذكر – ماهي	List	List	List of steps
هل	Did	Yes / No	Yes or No

Table 15. *Expected answer type identification*

```
Input - User Question in Natural Language

Output - List of words, Stemmed words, Keywords, Answer type, Focus detector

tokenList = tokenizer (userQuestion)

tokenWithoutStopWordsList = tokenList.stopWordsRemover ()

interrogativeNoun = tokenList.interrogativeNoun ()

if(interrogativeNoun == "اين")
        Then {answerExpectedType = Location}
else if(interrogativeNoun == "من" || interrogativeNoun == "لمن")
        Then {answerExpectedType = Person}
else if(interrogativeNoun == "متى")
        Then {answerExpectedType = Time}
else if(interrogativeNoun == "ما" || interrogativeNoun == "مما")
        Then {answerExpectedType = Object}
else if(interrogativeNoun == "كم")
        Then {answerExpectedType = Quantity}
else if(interrogativeNoun == "كيف")
        Then {answerExpectedType = Process}
else if(interrogativeNoun == "لماذا")
        Then {answerExpectedType = Reason}
else if(interrogativeNoun == "ماهي" || interrogativeNoun == "اذكر")
        Then {answerExpectedType = List}
else if(interrogativeNoun == "هل")
        Then {answerExpectedType = Yes/No}
else
        Then {answerExpectedType = Null}

questionKeywords = tokenList.questionKeywords()

stemmedQuestionKeywords = stemmer(questionKeywords)

extraKeywords = focusDetector.Generator(questionKeywords)
```

Figure 14. *Question analysis algorithm*

Table 16 illustrates the output of the question analysis component for the example of question "*متى تأسست الجامعة البريطانية في دبي؟*" (When was The British University in Dubai established?).

Task	Output
Tokenizer	متى – تأسست – الجامعة – البريطانية – في – دبي - ؟
Stop word removal	تأسست – الجامعة – البريطانية – دبي
Stemmer	تاسس – جامع – بريطان – دبي
Keywords extractor	تأسست
Class extractor	متى
Expected answer type	Time
Extra keywords	اقام – انشأ – بني – شيد هيئة – منظمة – جمعية – تجمع

Table 16. *Question analysis output*

4.5.2. Passage Retrieval

This component receives a list of question's terms as an input to retrieve the relevant passages from the corpus. The purpose of this component is not to extract the answers, but to identify the most relevant passages that may contain the question's answer.

The statistical approach is used to retrieve and rank the relevant passages. The keyword-based approach is utilized to retrieve the passages by calculating the degree of similarity between each passage and the question's terms (Keyword in actual and stem forms, as well as Extra Keywords). While, the Cosine Method is utilized to rank the retrieved passages, as well as Distance Density (N-gram) approach between keywords and retrieved passages. Accordingly, the top-ranked passages are picked as a candidate passage.

Indeed, the effectiveness of this component play a key role in the overall performance of QA system. If it fails to return the relevant passages that contains an answer, the next component will also inevitably fail to determine the correct answer. Figure 15 presents the high-level design algorithm for this component.

```
Input - Question type, List of stemmed question keywords, Extra keywords

Output - List of potential passages

n = 0           // Number of retrieved passages

while not endOfCorpus
{
   potentialAnswer [n] = locateMatchingSentence (questionkeywordsrds,
                                 stemmedQuestionKeywords, extraKeywords)

   n++

}

i = 0           // Counter for potentialAnswer list

while end of potentialAnswer
{
   passageCosineRank = potentialAnswer[i].CosineRank ()
   PassageDistanceDensityRank = potentialAnswer[i].DensityRank ()
   passageRankList.Add (potentialAnswer[i], passageCosineRank + PassageDistanceDensityRank)
   i++
}

passageRankList.sortDescending()
```

Figure 15. *Passage retrieval algorithm*

Table 17 illustrates the inputs required for the passage retrieval component, and the retrieved passages are ranked based on the best matches for the example of question

من هو الطبيب المتخصص بعمليات قسطرة القلب؟

Inputs Parameters	
Stemmer	طبيب – متخصص – عمليا – قسطر- قلب
Keywords extractor	القسطرة – القلب
Class extractor	من
Expected answer type	NE (Person)
Extra keywords	مطبب – معالج - مداو اختصاصي – انفرد به
Retrieved Passages	
Passage 1	طبيب اختصاصى قلب وشرايين :دراسة أعراض الشكوى المرضية والتاريخ المرضي له ودراسة التقارير والإجراءات الطبية السابقة للمريض. إجراء الفحص السريري الشامل وبخاصة القلب والشرايين، وقياس العلامات الحيوية اللازمة، وتحديد الفحوصات المخبرية والشعاعية وتخطيط

	القلب، وفحص الجهد والصور المقطعية والتشخيصية اللازمة، وقراءة وتحليل نتائج الفحوصات، وتشخيص الحالة المرضية. وضع خطة العلاج، وتحديد الإجراءات العلاجية اللازمة، وتحديد برنامج الرعاية اللازمة، ومتابعة الإجراءات العلاجية، وتقييم نتائجها. إجراء عملية القسطرة، وإجراء عمليات توسيع الشرايين بالبالون مع اختصاصي جراحة الأوعية الدموية وتحديد نوع الشبكة اللازم تركيبها للمريض، ووضع الشبكات اللازمة بحسب حالة المريض، وتحديد العلاجات اللازمة وكيفية استعمالها، وإعداد تعليمات التثقيف الصحي للمريض وأسرته، ومتابعة تنفيذها، وتحويل بعض الحالات المرضية التي تحتاج إلى تداخلات طبية من الاختصاصات الأخرى. حفظ وتوثيق البيانات المرضية والعلاجية للمريض. متابعة وتطوير أنظمة ومعايير الجودة في مجال أمراض القلب والشرايين.
Passage 2	طبيب اختصاصي أوعية دموية :دراسة الحالات المرضية المتعلقة بالأوعية الدموية سواء كانت وراثية أم مكتسبة، وإجراء الكشف السريري الدقيق والتفصيلي لهؤلاء المرضى، وطلب الفحوصات المخبرية والصور الشعاعية والتنظيرية والفحوصات المتخصصة بالأوعية الدموية مثل تخطيط الأوعية الدموية، وقراءة وتحليل نتائج الفحوصات والتقارير، والتنسيق مع الطبيب الاختصاصي في أمراض الدم واختصاصي جراحة الأوعية الدموية لتشخيص الحالة المرضية. إعداد خطة العلاج والرعاية اللازمة، ومتابعة تنفيذها، وإشراك الأهل باتخاذ القرار المتعلق بخطة الرعاية، وتوضيح كيفية استعمال العلاجات وسبل الرعاية المتاحة بما فيها مصلحة وصحة المريض، وتحديد العلاجات وكمياتها ومواعيد تناولها، والإشراف مع اختصاصي جراحة الأوعية الدموية عند إجراء عمليات توسيع الأوعية الدموية بالبالونات أو تغييرها، ومتابعة خطة العلاج، وتقييم النتائج. تحويل الحالات التي تتطلب رعاية متقدمة أو تدخلات جراحية أو تشخيصية لا تقع ضمن اختصاصه، وتقديم الاستشارات المتعلقة بالأوعية الدموية إلى الأطباء المعالجين الآخرين، ومتابعة وتقييم مدى فاعلية العلاج، وحفظ وتوثيق البيانات المرضية والعلاجية للمريض. متابعة وتطوير أنظمة ومعايير الجودة في مجال الأوعية الدموية.
Passage 3	طبيب اختصاصي نقل دم :أخذ السيرة الذاتية والمرضية للمتبرع، والتأكد من خلو المتبرع من الأمراض السارية والمعدية، وفحص الدم مخبريا للتأكد من خلوه من الأمراض، مثل التهاب الكبد والايدز وغيرها، وتحديد قوة وزمرة الدم، وإجراء فحوصات الأمراض المزمنة مثل السكري والضغط، وتوثيق العمليات الجراحية للمتبرع، وإجراء الفحص السريري للمتبرع، وقياس العلامات الحيوية من ضغط دم ونبض ودرجة حرارة وأخذ الوزن والطول، وضبط لياقة المتبرع للتبرع بالدم وتثقيف المتبرع والمجتمع بأهمية التبرع بالدم. تجهيز معدات سحب الدم، وتجهيز معدات ووحدات تعبئة الدم، والإشراف على سحب الدم، ومراقبة الحالة الصحية للمتبرع في أثناء

	السحب، وتحديد ردود الفعل عند المتبرع، **ومعالجتها**. مراقبة وحدات الدم، وفحص تجلطات الدم في الوحدة، وأخذ العينات اللازمة لفحص الوحدة مخبريا، وتقرير خلوها من الأمراض. فصل مكونات الدم مثل البلازما والصفائح الدموية، وإجراء **المعالجة** الكيماوية والفيزيائية للدم. تحديد أسس وأساليب حفظ وتخزين الدم، ومراقبة ثلاجة حفظ وحدات الدم وضبط درجة حرارتها وإتلاف الدم الفاسد، وإتلاف وحدات الدم التي انتهت صلاحيتها. دراسة طلبات المستشفيات من وحدات الدم، وصرف الدم بحسب الأصول. إعطاء الدم للمريض، وتقييم استجابة المريض للدم الجديد بالتنسيق مع الأطباء **الاختصاصين** والجراحين، وأخذ عينات من المريض ووحدة الدم، ومراقبة حدوث مضاعفات في أثناء نقل الدم إلى المريض. توثيق إجراءات سحب الدم، وتوثيق إجراءات إعطاء الدم، وتوثيق إجراءات إتلاف الدم. متابعة وتطوير أنظمة ومعايير الجودة في مجال نقل الدم.
Passage 4	طبيب **اختصاصي** أمراض الدم :دراسة الحالات المرضية المتعلقة بأمراض الدم والغدد اللمفاوية، وأخذ السيرة الذاتية والمرضية والأعراض والشكوى المرضية للحالة من المريض أو أهله، ودراسة التقارير الطبية السابقة المتعلقة بالمريض، وأخذ العلامات الحيوية اللازمة من قياس حرارة ونبض وقياس الضغط وغيرها. فحص المريض فحصا سريريا شاملا وكاملا، وتحديد الفحوصات المخبرية والكيمائية للدم، وتحديد الفحوصات المخبرية الأخرى من دم وكيمياء حيوية، وتحديد الفحوصات المجهرية للدم، وتحديد الفحوصات البيولوجية والجرثومية الدم، وأخذ خزعة النخاع الشوكي من القفص الصدري، وتحديد الفحوصات الشعاعية وفوق الصوتية والنووية اللازمة، وتشخيص الحالة المرضية من أمراض الدم أو سرطان الدم والجهاز اللمفاوي. إعداد خطة العلاج مع الفريق الطبي والتمريضي **والاختصاصات** الطبية الأخرى، وتحديد نوع العلاج الجراحي أو الكيماوي أو النووي، وتحويل الحالات التي تحتاج إلى تداخلات طبية من الاختصاصات الأخرى. متابعة ومراقبة تطورات الحالة الصحية للمريض، وتقييم فاعلية العلاج، وحفظ وتوثيق البيانات المرضية والعلاجية للمريض. متابعة وتطوير أنظمة ومعايير الجودة في مجال أمراض الدم.

Table 17. *Question retrieval output*

Besides, due to the identification of candidates for posted jobs is a time-consuming task for companies, as well as inappropriateness of traditional information retrieval techniques like the Boolean search methods. We can utilize this component as a recommender system by matching the resume skills with ACO job description in order to retrieve the most relevant jobs for posted resume.

4.5.3. Answer Extraction

This component receives relevant passages as an input to extract the most relevant answer, considering inputs from the Question Analysis component. The component performs its tasks differently based on the answer type that received from the Question Analysis component. It performs a complete parse of each received passages and identifies the candidate's answer by comparing it with the answer type. The components consist of three modules, namely: (i) Answer Stemmer, which fetches the root keywords in the retrieved passages; (ii) Similarity Checker, which facilitates the verification of keywords similarity between the question and the potential answers by measuring the number of matching keywords of the retrieved passages and the question; (iii) Answers Ranker, which sorts answers based on their score of similarity. **Error! Reference source not found.** presents the implemented algorithm of this component.

```
Input - List of potential passages, Stemmed question keywords

Output - List of extracted answers ranked in descending order of similarity

n = 0           // No. of potential answers

While number(List of potential passages)
{
    stemmedAnswerKeywords = stemmer(potentialAnswer[n])
    similarity = noOfSimilarWords(stemmedQuestionKeywords, stemmedAnswerkeywords)
    potentialAnswer[n].similarity = similarity
    sortDescending (potentialAnswer[n])
    n++
}
```

Figure 16. *Answer extraction algorithm*

Table 18 illustrates the inputs required for the answer extraction component, and the retrieved list of expected answer for the example of question

من هو الطبيب المتخصص بعمليات قسطرة القلب؟

Inputs Parameters	
Stemmer	طبيب – متخصص – عمليا – قسطر- قلب
Ranked passages	P1, P2, P3, P4
Expected answer type	NE
Output Parameters	
Potential answer ranked	1. طبيب اختصاصي قلب وشرايين (اسم وظيفة)

2. طبيب **اختصاصى** أوعية دموية 3. طبيب **اختصاصى** نقل دم 4. طبيب **اختصاصى** أمراض الدم	

Table 18. *Potential answer*

Based on the answer type that was received from the Question Analysis component, answer processing will be performed accordingly. For instance, using NER with patterns is a successful approach to choose the answers for factoid and definitional questions. Whereas, the complex questions (why/how) require a semantic parsing to extract answers.

The potential answers of why questions have common phrases that are considered as unit connectors, such as: (i) connecter words for result relation (لذا - نتيجة لذلك - نتيجة لهذا - نســتنتج); (ii) connecter words for justification relation (تعليل ذلك - بسبب - لان). In view of the above, the example of the questions below and Table 19 illustrate how to deal with this kind of question.

كيف يمكن الوصول الى الاسترخاء العميق؟

ماذا يحدث في حالة تنفس متعادل بين الشهيق والزفير؟

Inputs Parameters	
Ranked passages	يمكنك الوصول إلى حالة من الاسترخاء العميق **من خلال** تنفس متعادل، يكون فيه كل شهيق وكل زفير طويلين ويساوي كل منهما الآخر في الطول.
Unit 1 of passage	يمكنك الوصول إلى حالة من الاسترخاء العميق
Relation type	Result relation
Unit 2 of passage	من خلال تنفس متعادل، يكون فيه كل شهيق وكل زفير طويلين ويساوي كل منهما الآخر في الطول

Table 19. *Why question example*

According to the similarity checker function, the first question matches with the first unit, which means that the answer is the second unit. Also, the second question is matched with the second unit, which means that the answer is the first unit.

4.6. Implementation

The system is a web-based Arabic QA system built from scratch using Asp.Net and C# within a DotNet framework version 4.6. It has a simple interface where the user can post a question

in the Arabic language. The user's question must begin with an interrogative noun. Otherwise, it is considered not a question and gives an error message.

Figure 17 shows the home page of the system. It is noted that the user must choose one corpus from the list of corpora to be used as a data source. The purpose of using more than one corpus is to evaluate the system performance based on Gold Standard dataset.

Figure 17. *Our Arabic QA system*

4.7. Evaluation Results

Evaluation is one of the most important pillars in this field, which assesses the performance of QA systems, as well as benchmarks to demonstrate how far the development has been achieved in this field. Thus, the experimental results of our participation in the IR field are evaluated. The evaluation process is carried out through a typical mechanism of four stages as shown in Figure 18. The first two stages, the Training and Dry-run are executed to adapt our system and make any necessary adjustments. While the last two stages, the Actual Running and Benchmarking are performed to figure out the system's outcomes based on gold-standard data and compare them with other well-established Arabic QA systems in order to obtain the ranking.

Figure 18. *QA Evaluation process*

4.7.1. Performance Measures

Generally, there is a range of tests required to calculate the effectiveness of the QA system. This part aims to present the measurement formulas that will be used to assess the efficiency of our system. Table 20 displays the formulas and their descriptions along with equations (Ray & Shaalan 2016).

Formula	Description	Equation
Precision (P)	Represents the percentage of retrieved documents related to the query	$P = \frac{Number\ of\ relevant\ documents}{\text{Retrieved documents}}$
Recall (R)	Represents the percentage of related documents retrieved	$R = \frac{Number\ of\ relevant\ documents}{\text{Relevant documents}}$
F - Measure (F)	Represents the percentage of combination for precision and recall	$F = \frac{2*P*R}{\mathrm{P} + \mathrm{R}}$

Table 20. *QA measurement formulas*

4.7.2. Results and Discussion

This part depicts the results of our system experiments that were conducted to measure the system performance in the context of QA. The evaluation methods are applied to factoid and complex questions based on the CLEF and TREC tracks (Open-domain corpus), as well as ACO (Close-domain corpus). The data sets are divided into training and testing sets in order to train our QA modules and then apply these modules to the test set.

We selected a set of 100 CLEF's questions and 100 TREC's questions that translated into Arabic language, as well as define 30 questions for ACO in order to evaluate the system performance. Table 21 details the results obtained for different data sets. It provides information

about the number of questions in each data set along with the calculated percentage of Recall, Precision and F-Measure.

Dataset	# Question	# Retrieved	# Answered	Precision	Recall	F-Measure
CLEF Evaluation (Factoid)	50	46	21	46%	92%	61%
CLEF Evaluation (Complex)	50	44	11	25%	88%	39%
TREC Evaluation (Factoid)	50	44	19	43%	88%	58%
TREC Evaluation (Complex)	50	41	9	22%	82%	35%
ACO Evaluation (Factoid)	20	18	10	56%	90%	69%
ACO Evaluation (Complex)	10	8	2	25%	80%	38%

Table 21. *Detailed experiments results*

Obviously, the obtained performance of our system showed that the highest scores were in the Factoid questions. While the complex questions have low scores, because this type of questions looking at descriptive answers that require more advanced techniques. As illustrated in Table 22, the system performance shows an average precision of 36%, by answering 72 questions out of 230 questions.

Type	Percentage
Precision	36%
Recall	87%
F-Measure	51%

Table 22.*Average evaluation results*

4.8. Chapter Summary

In this chapter, we have highlighted the motivation and aim that attracted us to build our QA system. Besides, we have detailed the methodology cycle which compose of 4 stages, namely: (i) data preparation; (ii) design the system algorithm; (iii) develop the system and conduct the necessary experiments; (iv) evaluate the effectiveness of the system. Moreover, we have explained our system structure which composed of three major components, namely: Question Analysis, Passage Retrieval and Answer Extraction. While each component consists of several modules with a distinct function. Moreover, we have produced an appropriate web application that help us select a data source and ask the question in order to answer based on the selected data source.

Accordingly, we explained the conducted experiments on a set of 230 question from TREC, CLEF, and ACO corpus. Obviously, the performance of our system showed that the highest scores

were in the Factoid questions. While the complex questions have low scores, because this type of questions looking at descriptive answers that require more advanced techniques. The system performance shows an average precision of 36%, by answering 72 questions, as well as the Recall was 78% and F-Measure was 51%.

Chapter Five: Research Question Answers

According to previous chapters, the results obtained were positively answered the research questions as follows:

- Is it possible to build a QA system that can answer different types of Arabic questions (Factoid and Complex)?

  Chapter four explains our system structure which composed of three major components, namely: Question Analysis, Passage Retrieval and Answer Extraction. While each component consists of several modules with a distinct function. Moreover, this chapter illustrates the web application that help us select a data source and ask the question (Factoid, Complex) in order to answer based on the selected data source.

- Is it possible to achieve acceptable performance even with different data source?

  Chapter four explains the experiments conducted on a set of 230 question from TREC, CLEF, and ACO corpus. The system performance shows an average precision of 36%, by answering 72 questions, as well as the Recall was 78% and F-Measure was 51%.

- Is it possible to build an Arabic annotated corpus to be used as one of data source?

  Chapter three produced the ACO corpus of one million words written MSA format. The corpus contains 700 occupations which are analyzed carefully and manually annotated.

Chapter Six: Conclusion and Future Work

This chapter presents a conclusion of what we have done in this thesis, as well as highlights the future work

6.1. Conclusion

In this thesis, we highlighted the importance of the linguistic resource to the IR systems and the aim that attracted us to build the ACO corpus. Also, we have produced an appropriate application to help us explore and tag (annotate) the corpus content. The corpus contains 700 occupations which are analyzed carefully and manually annotated. Moreover, we used Cohen's Kappa coefficient method to evaluate the reliability of the tagged content. The corpus content has been tagged and assessed by two different groups of taggers. Moreover, the synonyms are added to the name of the job itself, as well as the context in which the name of the job or the nature of the work may come. Accordingly, the inter-annotator agreement indicates that the reliability of ACO corpus is almost perfect agreement. As well as, the content of the corpus is highly confidence and reliable according to the result achieved by 90%.

Besides, we explained the structure of our QA system which composed of three major components, namely: Question Analysis, Passage Retrieval and Answer Extraction. While each component consists of several modules with a distinct function. Also, we have produced an appropriate web application that help us select a data source and ask the question in order to answer based on the selected data source. Accordingly, we conducted experiments on a set of 230 question from TREC, CLEF, and ACO corpus. The system performance shows an average precision of 36%, by answering 72 questions, as well as the Recall was 78% and F-Measure was 51%.

6.2. Future Prospects

Over the past ten years, linguistic research has been reliant on text corpora to represent the Arabic language. Currently, the available reliable Arabic corpora were not sufficient for IR research. There is a lot of future work to improve the ACO, such as adapting the mining data approach by launching an information extraction tool, as well as improving the PoS-tag performance by incorporating more features.

Besides, our Arabic QA system has shown promising performance, we plan in the future to improve the question classifier using patterns, as well as enhance the answer extraction of the complex question and answer validation using web.

References

Albarghothi, A., Khater, F. & Shaalan, K. (2017). Arabic Question Answering Using Ontology. *Procedia Computer Science*, vol. 117, pp. 183-191.

Salloum, S., AlHamad, A., Al-Emran, M. & Shaalan, K. (2018). A survey of Arabic text mining. *Intelligent Natural Language Processing*, pp. 417-431.

Shaalan, K., Al-Sheikh, S. & Oroumchian, F. (2012). Query expansion based-on similarity of terms for improving Arabic information retrieval. *International Conference on Intelligent Information Processing*, pp. 167-176.

Al-Chalabi, H., Ray, S. & Shaalan, K. (2015). Semantic Based Query Expansion for Arabic Question Answering Systems. *Arabic Computational Linguistics* (ACLing), pp. 127-132.

Cheddadi, A. (2014). Three-levels Approach for Arabic Question Answering Systems. *Diss. Ecole Mohammadia d'Ingénieurs.*

Shaalan, K. (2010). Rule-based approach in Arabic natural language processing. *The International Journal on Information and Communication Technologies (IJICT)*, pp. 11-19.

Green, B., Wolf, A., Chomsky, C. & Laughery, K. (1961). Baseball: an automatic question-answerer. *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference on - IRE-AIEE-ACM '61 (Western).*

Woods, W. (1973). Progress in natural language understanding: an application to lunar geology. *Proceedings of the June 4-8, 1973, national computer conference and exposition on - AFIPS '73.*

Clark, P., Thompson, J. & Porter, B. (1999). A knowledge-based approach to question-answering. *Proc. AAAI*, vol. 99, pp. 43-51.

Al-Sughaiyer, I. & Al-Kharashi, I. (2004). Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology*, vol. 55 (3), pp. 189-213.

Kwok, C., Etzioni, O. & Weld, D. (2001). Scaling question answering to the web. *ACM Transactions on Information Systems*, vol. 19 (3), pp. 242-262.

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

Al Chalabi, H., Ray, S. & Shaalan, K. (2015). Question classification for Arabic question answering systems. *Information and Communication Technology Research (ICTRC)*, pp. 310-313.

Shaalan, K. & Raza, H. (2009). NERA: Named entity recognition for Arabic. *Journal of the American Society for Information Science and Technology*, pp. 1652-1663.

Rahman, T. (2015). Question classification using statistical approach: a complete review. *Journal of Theoretical and Applied Information Technology*, pp. 386-395.

Bronner, A. & Monz, C. (2012). User edits classification using document revision histories. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 356-366.

Rosso, P., Lyhyaoui, A., Peñarrubia, J., Montes, M., Gómez, M., Benajiba, Y. & Raissouni, N. (2005). Arabic-english question answering. *Proceeding of Information Communication Technologies Int. Symposium, ICTIS-2005, Tetuan, Morroco, June*, pp. 3-6.

Kanaan, G., Hammouri, A., Al-Shalabi, R. & Swalha, M. (2009). A new question answering system for the Arabic language. *American Journal of Applied Sciences*, vol. 6, pp. 797-805.

Jurafsky, D. & Martin, J. (2014). *Speech and language processing*. London: Pearson.

Navarro, G., Puglisi, S. & Valenzuela, D. (2015). General document retrieval in compact space. *Journal of Experimental Algorithmics (JEA)*, pp. 2-3.

Prager, J. (2007). Open-domain question–answering. *Foundations and Trends in Information Retrieval*, pp. 91-231.

Corston, S., Dolan, W., Vanderwende, L. & Braden-Harder, L. (2005). System for processing textual inputs using natural language processing techniques. *U.S. Patent 6,901,399*.

Benajiba, Y. & Rosso, P. (2007). Arabic Question Answering. *Diploma of advanced studies. Technical University of Valencia, Spain*.

Abouenour, L. (2011). On the Improvement of Passage Retrieval in Arabic Question/Answering (Q/A) Systems. *Natural Language Processing and Information Systems*, pp. 336-341.

Toba, H., Ming, Z., Adriani, M. & Chua, T. (2014). Discovering high quality answers in community question answering archives using a hierarchy of classifiers. *Information Sciences*, pp. 101-115.

Trigui, O., Belguith, L. & Rosso, P. (2010). DefArabicQA: Arabic definition question answering system. *Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC, Valletta, Malta*, pp. 40-45.

Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Goodrum, R., Girju, R. & Rus, V. (1999). Lasso: A tool for surfing the answer net. *TREC*, vol. 8, pp. 65-73.

Harabagiu, S., Moldovan, D., Pasca, M., Surdeanu, M., Mihalcea, R., Girju, R., Rus, V., Lacatusu, V., Morarescu, P. & Bunescu, R. (2001). Answering Complex, List and Context Questions with LCC's Question-Answering Server. *TREC*.

Srihari, R. & Li, W. (1999). Information extraction supported question answering. *CYMFONY NET INC WILLIAMSVILLE NY*.

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

Clarke, C., Cormack, G., Lynam, G., Terra, E. & Tilker, P. (2002). Statistical Selection of Exact Answers (MultiText experiments for TREC 2002). *NIST Publication*, pp. 162-170.

Khalid, M. & Verberne, S. (2008). Passage retrieval for question answering using sliding windows. *Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, pp. 26-33.

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*, vol. 21 (2), pp. 133-154.

LLopis, F., Vicedo, J. & Ferrández, A. (2002). Passage selection to improve Question Answering. *proceeding of the 2002 conference on multilingual summarization and question answering - COLING-02*, pp. 1-6.

Harman, D. (1992). The DARPA TIPSTER project. *ACM SIGIR Forum*, vol. 26 (2), pp. 26-28.

Pallett, D. (n.d.). A look at NIST'S benchmark ASR tests: past, present, and future. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*.

Ray, S. & Shaalan, K. (2016). A Review and Future Perspectives of Arabic Question Answering Systems. *IEEE Transactions on Knowledge and Data Engineering*, vol. 28 (12), pp. 3169-3190.

Voorhees, E. & Harman, D. (2005). TREC experiment and evaluation in information retrieval. *Cambridge, Mass.:MIT Press*.

EDMONDS, P. & KILGARRIFF, A. (2002). Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, vol. 8 (4), pp. 279-291.

Hauser, R. (1994). Results of the 1. *Morpholympics, LDV-FORUM*, vol. 11.

Adda, G., Mariani, J., Lecomte, J., Paroubek, P. & Rajman, M. (1998). The GRACE French part-ofspeech tagging evaluation task. *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, vol. 1, pp. 433-441.

Agosti, M., Di Nunzio, G., Ferro, N., Harman, D. & Peters, C. (2007). The Future of Large-scale Evaluation Campaigns for Information Retrieval in Europe. *International Conference on Theory and Practice of Digital Libraries*, pp. 509-512.

Mitkov, R. (2005). The Oxford handbook of computational linguistics. *Oxford Handbooks in Linguistics, Oxford University Press*.

Farghaly, A. & Shaalan, K. (2009). Arabic Natural Language Processing. *ACM Transactions on Asian Language Information Processing*, vol. 8 (4), pp. 1-22.

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O. & Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *LREC*, vol. 14, pp. 1094-1101.

Abdelali, A., Darwish, K., Durrani, N. & Mubarak, H. (2016). Farasa: A fast and furious segmenter for arabic. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 11-16.

Shaalan, K., Rafea, A., Abdel Monem, A. & Baraka, H. (2004). Machine Translation of English Noun Phrases into Arabic. *International Journal of Computer Processing of Languages*, vol. 17 (02), pp. 121-134.

Buckwalter, T. (2004). Buckwalter Arabic Morphological Analyzer Version 2.0. *Linguistics Data Consortium (LDC).*

Harmanani, H., Keirouz, W. & Raheel, S. (2006). A rule-based extensible stemmer for information retrieval with application to Arabic. *International Arab Journal of Information Technology*, vol. 3.

Vergyri, D. & Kirchhoff, K. (2004). Automatic diacritization of Arabic for acoustic modeling in speech recognition. *Proceedings of the workshop on computational approaches to Arabic script-based languages*, pp. 66-73.

Habash, N. & Rambow, O. (2007). Arabic diacritization through full morphological tagging. *The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp. 53-56.

Farghaly, A. & Shaalan, K. (2009). Arabic Natural Language Processing. *ACM Transactions on Asian Language Information Processing*, vol. 8 (4), pp. 1-22.

Shaalan, K. (2014). A Survey of Arabic Named Entity Recognition and Classification. *Computational Linguistics*, vol. 40 (2), pp. 469-510.

Abouenour, L., Bouzoubaa, K. & Rosso, P. (2013). Erratum to: On the evaluation and improvement of Arabic WordNet coverage and usability. *Language Resources and Evaluation*, vol. 47 (4), pp. 1343-1343.

Shaheen, M. & Ezzeldin, A. (2014). Arabic Question Answering: Systems, Resources, Tools, and Future Trends. *Arabian Journal for Science and Engineering*, vol. 39 (6), pp. 4541-4564.

Benajiba, Y., Rosso, P. & Soriano, J. (2007). Adapting the JIRS passage retrieval system to the Arabic language. *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 530-541.

Trigui, O., Belguith, L. & Rosso, P. (2010). DefArabicQA: Arabic definition question answering system. *Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC, Valletta, Malta*, pp. 40-45.

Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Forascu, C. & Sporleder, C. (2011). Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation. *CLEF*, pp. 1-20.

Oudah, M. & Shaalan, K. (2016). Studying the impact of language-independent and language-specific features on hybrid Arabic Person name recognition. *Language Resources and Evaluation*, vol. 51 (2), pp. 351-378.

Dwivedi, S. & Singh, V. (2013). Research and Reviews in Question Answering System. *Procedia Technology*, vol. 10, pp. 417-424.

Riloff, E. & Thelen, M. (2000). A rule-based question answering system for reading comprehension tests. *ANLP/NAACL 2000 Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems*, vol. 6, pp. 13-19.

Hao, X., Chang, X. & Liu, K. (2007). A Rule-based Chinese Question Answering System for Reading Comprehension Tests. *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007)*, vol. 2, pp. 325-329.

Ittycheriah, A., Franz, M., Zhu, W., Ratnaparkhi, A. & Mammone, R. (2000). IBM's Statistical Question Answering System. *Proceedings of the Text Retrieval Conference TREC-9*.

Berger, A., Caruana, R., Cohn, D., Freitag, D. & Mittal, V. (2000). Bridging the lexical chasm. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '00*, pp. 192- 199.

Al-Shawakfa, E. (2016). A Rule-Based Approach to Understand Questions in Arabic Question Answering. *Jordanian Journal of Computers and Information Technology (JJCIT)*, vol. 2.

Ravichandran, D. & Hovy, E. (2001). Learning surface text patterns for a Question Answering system. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pp. 41-47.

Zhang, D. & Lee, W. (2002). Web Based Pattern Mining and Matching Approach to Question Answering. *Proceedings of the 11th Text REtrieval Conference*, vol. 2, p. 497.

Gunawardena, T., Lokuhetti, M., Pathirana, N., Ragel, R. & Deegalla, S. (2010). An automatic answering system with template matching for natural language questions. *2010 Fifth International Conference on Information and Automation for Sustainability*, pp. 353-358.

Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A., Gerber, D. & Cimiano, P. (2012). Template-based question answering over RDF data. *Proceedings of the 21st international conference on World Wide Web - WWW '12*, pp. 639-648.

Nwesri, A. (2008). Effective retrieval techniques for Arabic text. *School of Computer Science and Information Technology, RMIT University*.

Mohammed, F., Nasser, K. & Harb, H. (1993). A knowledge based Arabic question answering system (AQAS). *ACM SIGART Bulletin*, vol. 4 (4), pp. 21-30.

Hammo, B., Abu-Salem, H. & Lytinen, S. (2002). QARAB: A question answering system to support the Arabic language. *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, pp. 1-11.

Benajiba, Y., Rosso, P. & BenedíRuiz, J. (2007). Anersys: An arabic named entity recognition system based on maximum entropy. *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 143-153.

Brini, W., Ellouze, M., Trigui, O., Mesfar, S., Belguith, L. & Rosso, P. (2009). Factoid and definitional Arabic question answering system. *Post-proceedings of NOOJ-2009*, pp. 8-10.

Bekhti, S., Rehman, A., Al-Harbi, M. & Saba, T. (2011). AQUASYS: AN ARABIC QUESTION-ANSWERING SYSTEM BASED ON EXTENSIVE QUESTION ANALYSIS AND ANSWER RELEVANCE SCORING. *International Journal of Academic Research*, vol. 3, pp. 45-54.

Trigui, O., Belguith, L., Ben Amor, H. & Gafsaoui, B. (2012). Arabic QA4MRE at CLEF 2012: Arabic Question Answering for Machine Reading Evaluation. *CLEF*.

Abdelnasser, H., Mohamed, R., Ragab, M., Mohamed, A., Farouk, B., El-Makky, N. & Torki, M. (2014). Al-Bayan: An Arabic Question Answering System for the Holy Quran. *Proceedings of the 9th International Workshop on Semantic Evaluation*, pp. 57–64.

Kurdi, H., Alkhaider, S. & Alfaifi, N. (2014). Developemnt and Evaluation of A Web Based Question Answering System for Arabic Languag. *International Journal on Natural Language Computing*, vol. 3 (2), pp. 11-32.

Nicosia, M., Filice, S., Barrón-Cedeno, A., Saleh, I., Gao, W., Mubarak, H., Nakov, P., Da San Martino, G., Moschitti, A., Darwish, K., Marquez, L., Joty, S. & Magdy, W. (2015). QCRI: Answer Selection for Community Question Answering Experiments for Arabic and English. *Proceedings of the 9th International Workshop on Semantic Evaluation*, pp. 203–209.

AL-Khawaldeh, F. (2015). Answer Extraction for Why Arabic Questions Answering Systems: EWAQ. *World of Computer Science and Information Technology Journal (WCSIT)*, vol. 5 (5), pp. 82-86.

Azmi, A. & Alshenaifi, N. (2017). Lemaza : An Arabic why-question answering system. *Natural Language Engineering*, vol. 23 (06), pp. 877-903.

Hamoud, B. & Atwell, E. (2017). Evaluation corpus for restricted-domain question-answering systems for the holy Quran. *International Journal of Science and Research*, vol. 6 (8), pp. 1133-1138.

Goweder, A. & De Roeck, A. (2001). Assessment of a significant Arabic corpus. *Arabic NLP Workshop at ACL/EACL*.

ISCO - International Standard Classification of Occupations". (2018). Available at: http://www.ilo.org/public/english/bureau/stat/isco/

Takala, P., Malo, P., Sinha, A. & Ahlgren, O. (2014). Gold-standard for Topic-specific Sentiment Analysis of Economic Texts. *LREC*, vol. 2014, pp. 2152-2157.

Kennedy, G. (2014). An introduction to corpus linguistics. Abingdon, *Oxon:Routledge*.

Alansary, S., Nagi, M. & Adly, N. (2007). Building an International Corpus of Arabic (ICA): progress of compilation stage. *7th international conference on language engineering*, pp. 5-6.