Sudan University of Science and Technology

College of Graduate Studies

A Question Answering System Design about the Holy Quran

تصميم نظام إجابة الأسئلة عن القرآن الكريم

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

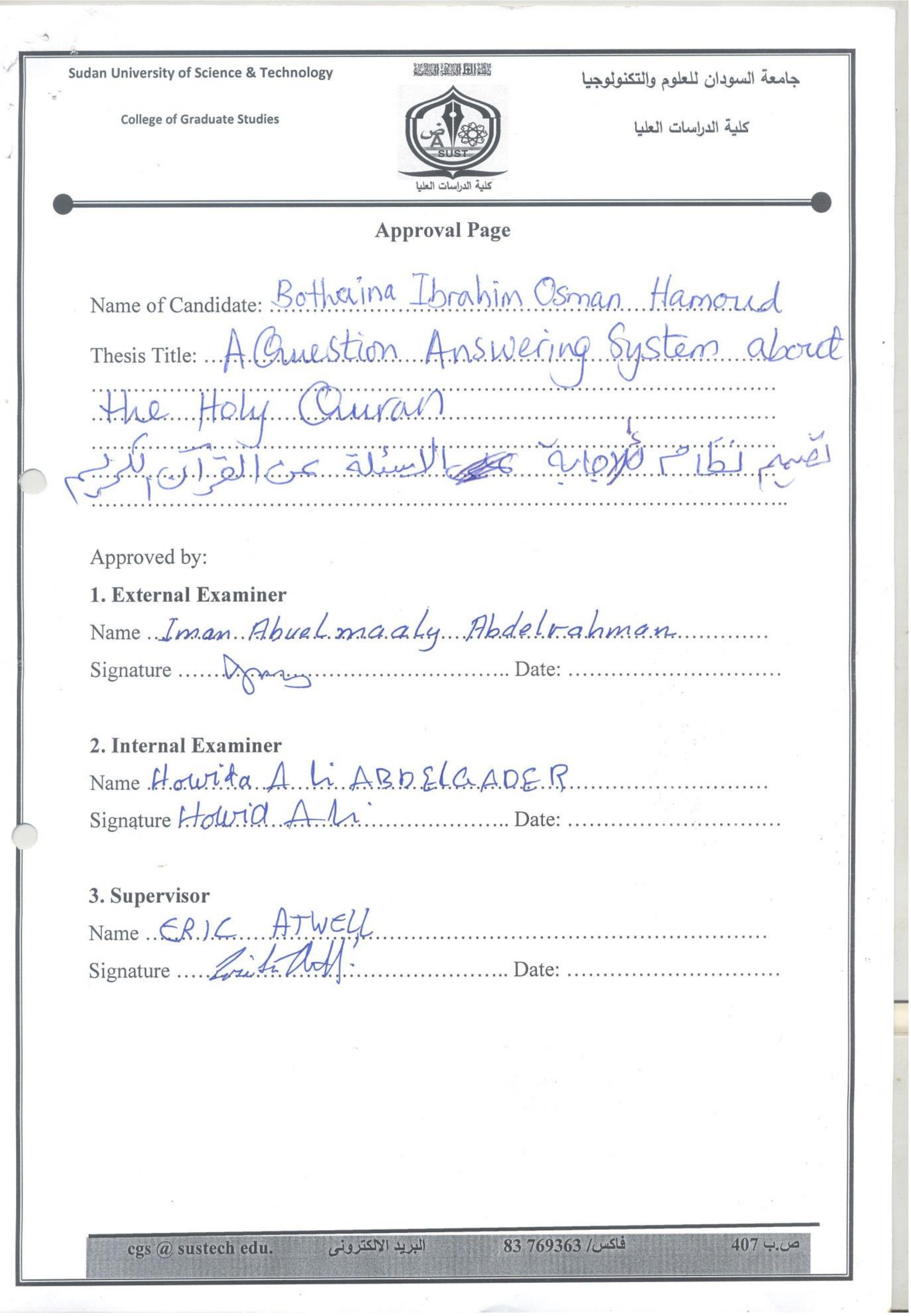
By:

Bothaina Ibrahim OsmanHamoud

Supervisor
Prof. Dr. Eric Atwell

August 2017

Approval page



Sudan University of Science & Technology

جامعة السودان للعلوم والتكنولوجيا

College of Graduate Studies

كلية الدراسات العليا

كلية الدراسات العليا

Approval Page

Name of Candidate: Bothaina Ibrahim Osman Hamoud

Thesis Title: A Question Answering System about the Holy Quran

تصميم نظام للإجابة الاسئلة عن القرآن الكريم

Approved by:

1. External Examiner

Name Iman Abuel maaly Abdelrahman

Signature .... Date: ....

2. Internal Examiner

Name Howita Ali ABDELGADER

Signature Howid Ali .... Date: ....

3. Supervisor

Name ERIC ATWELL

Signature .... Date: ....

cgs @ sustech edu. البريد الالكترونى 83 769363 /فاكس ص.ب 407

Sudan University of Science and Technology

College of Graduate Studies

A Question Answering System Design about the Holy Quran

تصميم نظام إجابة الأسئلة عن القرآن الكريم

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy inComputer Science

By:

Bothaina Ibrahim OsmanHamoud

Supervisor
Prof. Dr. Eric Atwell

August 2017

INTRODUCTIVE PAGE

"وَعَلَّمَ آدَمَ الْأَسْمَاءَ كُلَّهَا ثُمَّ عَرَضَهُمْ عَلَى الْمَلَائِكَةِ فَقَالَ أَنبِئُونِي بِأَسْمَاءِ هَٰؤُلَاءِ إِن كُنتُمْ
صَادِقِينَ(31).قَالُوا سُبْحَانَكَ لَا عِلْمَ لَنَا إِلَّا مَا عَلَّمْتَنَا إِنَّكَ أَنتَ الْعَلِيمُ الْحَكِيمُ (32)"

صدق الله العظيم

الآيتان (31)، (32) سورة البقرة

DEDICATION

To the soul of my mother, my beloved father, lovely husband, my supervisor for his continual support, sisters, brothers, friends, colleagues and to the whole Islamic nation

ACKNOWLEDGEMENTS

First and foremost I praise God (Allah) - for his generosity, mercy and bounty - who provided me the health, strength and patience, guiding me through all the difficulties to start and then complete this work. I would never be able to finish it without the will of Allah (swt). Many thank my father for your prayers and encouragement. Godbless you and forgive you.

I would like to express my sincere gratitude and thanks to my superb supervisor Dr. Eric Atwell at the University of Leeds. Many thanks for your regular excellent supervision, caring, and the considerable advices you provided to me, patiently listened to me in the Saturday Skype meetings, and gave me the best suggestions. Thank you for your endless help, support and guidance. I owe my deepest appreciation to the distinguished professors in Sudan University of Science and technology.

Special thanks to my dear’s lovely sisters Dr. Fatima, Shadia and Ekhlass, who had been with me at each moment encouraging supporting and helping.I am indebted to the deep understanding of my colleagues and friends in the deanship of Preparatory year/ Umm Al-Qura University, thanks a lot for your cooperation, encouragement, and help.

I would like also to thank all my colleagues in batch4 for sharing ideas, and working together during the first and second semester. Finally, I will be forever grateful for the kindness and support provided by my husband Ahmed Mahmoud. I thank you for your patience and continued support at every step of my way.

For all my beloved and even those whom I have not acknowledged, God bless you and grant you the health, happiness, and success in this world and the hereafter.

ABSTRACT

This research contributes to the area of Question Answering (QA) to develop a novel QA system for the holy Quran. Q A is an important research area concerned with developing an automated approach that answers questions posed by humans in a natural language. None of the existing QA systems on the Quran has been tested on a Quran question and answer corpus using question redundancy with the aim of question answering. In this thesis, a basic beginning arises from the usability and limitation of a restricted domain knowledge base. We considered the special domain knowledge base important and useful to answer new question asked by a user, as it contains a lot of similar or relevant question information. This work aims to compile a question and answer bilingual corpus in order to develop a QA system that answer Arabic and English questionsabout the holy Quran. This corpus has been collected from several credible sources,and hence it canbe used for the evaluation of question answering systems or any other application where questions and answers are needed. Considering these, first we investigated different techniques and tools necessary to build a novel QA system. WEKA and Nooj have been tested and explored by conducting some experiment, but they are not suited to build a QA system. Then we built our own implementation: QAEQAS a Quranic Arabic/English Question Answering System. As a complete QA solution, we used python and its toolkit to process the user question and the corpus, as well as to implement the search engine to retrieve candidate results and then extract the best answer. A central argument of this thesis is that it relies on a specialized search corpus, and data redundancy by integrating a range of knowledge bases from different sources as well as reformulating the corpus questions in different ways in differing context. Two prototype versions of QA system were developed. QAEQAS presents its usefulness that deals with a wide range of question types in addition to its high accuracy. QAEQAS used the most common evaluation metrics in Information Retrieval (IR) and QA to evaluate the effectiveness and correctness of the system results, namely precision and recall. The resultshave shown that the performance of the system is increased with more data redundancy;it registered a precisionand recall of 96% and 94%for Arabic, 90% and 89%for English respectively.

المستخلص

يساهم هذا البحث في مجال الإجابة على الأسئلة لتطوير نظام جديد للقرآن الكريم يجيب على أسئلة المستخدم. الإجابة علىالأسئلة (QA) هو مجال بحث هام يتعلق بوضع نهج آلي يجيب على الأسئلة التي يطرحها الإنسان باللغة الطبيعية. الأنظمة الموجودة للرد على أسئلة القرآن لم يتم اختبارها باستخدام مجاميع (corpus)تتكون من سؤال مع إجابته الصحيحة و باستخدام تكرار السؤال. تنشأ البداية الأساسية من استخدام والحد من قاعدة معرفة لمجال محدود. لقد اعتبرنا قاعدة المعرفة الخاصة بالمجال و التي تتكون من سؤال مع إجابته الصحيحة هامة ومفيدة للإجابة على سؤال جديد يطرحه المستخدم، لأنهاتحتوي على الكثير من معلومات الأسئلة المشابهة أو ذات الصلة. يهدف هذا العمل إلى تجميع أسئلة مع اجاباتها الصحيحة من اجل استخدامها في تطوير نظام جديديجيب على أسئلة القرآن العربيةوالإنجليزية. لقدتم جمع هذه المجاميع (corpus)منعدةمصادرموثوق بها،ومنثميمكناستخدامهالتقييمنظامالإجابةعلىالأسئلةأوأيتطبيقآخريحتاجإلىأسئلةوأجوبة القرآن الكريم. وبالنظر إلى هذا فقد تم اختبار واستكشاف الأدواتويكا (WEKA) ونووج(Nooj) من خلال إجراء بعض التجارب، ولكنها ليست مناسبة لبناء نظام الرد على أسئلة القرآن الكريم. بعد ذلك قمنابتطويرالتطبيقالخاصبنا: QAEQAS نظام الرد على أسئلة القرآن العربية والانجليزية. كحلكامللنظام الإجابةعلىالأسئلةاستخدمنا لغة بايثون(Python)ومجموعة ادواتها للغة الطبيعية(nltk)لمعالجة أسئلة المستخدم و المجاميع(corpus)، ثم استخدام محرك البحث لاسترجاع الأسئلة المرشحة ومن ثم استخراج أفضل إجابة. الحجة الاساسية لهذه الأطروحة هي أنها تعتمد فيالبحث على مجاميع(corpus) خاصةوتكرار البيانات من خلال دمج مجموعة من قواعد المعرفة من مصادر مختلفة، فضلا عن إعادة صياغة الأسئلة بطرق مختلفة في سياق مختلف. لقد تم تطوير نسختين من النموذج الأولي، و قد برهن النظام فائدته التي تتناول مجموعة واسعة من أنواع الأسئلة بالإضافةإلىدقةعالية. استخدم QAEQAS مقاييس التقييم الأكثر شيوعا والمستخدمة في استرجاع المعلومات والرد على الأسئلة لتقييم فعالية وصحة نتائج النظام، وهي الدقة والاسترجاع. وقد أظهرت النتائج أن أداء النظام يزداد معزيادة تكرار البيانات، لقد سجلدقة و استرجاع96% و 94%للغةالعربيةو90% و 89%للغةالإنجليزيةعلىالتوالي.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

AI	Artificial Intelligence
CQA	Community Question Answering
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma-Separated Values
DM	Data Mining
FAQ	Frequently Asked Questions
GUI	Graphical UserInterface
IE	Information Extraction
IR	Information Retrieval
KD/D	Knowledge-Discovery in Database
KDD	Knowledge Discovery in Databases
KDT	Knowledge-Discovery in Text
NER	Named Entity Recognition
NLP	Natural Language Processing
nltk	natural language toolkit
Q&A	Question and Answer
QA	Question Answering
QAEQAS	Quranic Arabic English Question Answering System
SVM	Support Vector Machine
SW	Semantic Web
WEKA	Waikato Environment for Knowledge Analysis

CHAPTER I

INTRODUCTION

1.1 General Overview

Advances in information technology and the tremendous amounts of electronic data on the internet give rise to the development of advanced search engines and document retrieval systems such as Google, AltaVista. A natural language search engine would in theory find targeted answers to user questions as opposed to keyword search. Conventional search engines ignore the question and instead search on the keywords. Natural language search, on the other hand, attempts to use natural language processing to understand the nature of the question and then to search and return a ranked list of short answers, or subset of the web that contains the answer to the question effectively, results would have a higher relevance than results from a keyword search engine. However in both cases the user currently bears the burden of searching through the returned result in order to find the required answer. That encouraged researchers to create question answering (QA) system which,aims at providing precise answer to question.

Automated Question Answering is the task of providing answers automatically to questions asked in natural language. Question answering system is a specialized form of Information Retrieval system, in which a direct answer is expected in response to a submitted query, rather than a set of references that may contain the answers. The fundamental aim behind the QA system is to facilitate human-machine interactions (Shawar and

Atwell, 2009). QA is concerned with developing an automated process that answers questions posed by humans in a natural language. A QA implementation, usually a computer program, may construct its answers by querying a structured database of knowledge or information, usually a knowledge base. QA systems can extract answers from an unstructured collection of natural language documents. The processing of a QA system may broadly compose of three stages, i.e., question analysis: parsing, question classification and query reformulation; document analysis: extract candidate documents, identify answers; and answer analysis: extract candidate answers and rank the best one.

Research on QA Systems dating back to sixties, in late 1990s the research communities decided to focus on factoid questions whose answers are relatively short and well-defined. Many different techniques and approaches have been investigated by many researchers during the last years in order to provide reliable QA systems for answering online user questions. These recent trends in QA have led to numerous studies focusing on presenting answers in a form which closely resembles a human answer.

1.2 Problem Statement and its Significance

Many people do not find accurate answers for their Islamic questions although the Quran provides wide and thorough teaching, not only about rituals but also about all aspects of life.Many people prefer to find their religious answers from the internet. Most search methods available through search engines cannot satisfy all users' needs to get exact and specific information. The user bears the burden of searching through all the documents(Abdelnasser et. al, 2014) to find the required answer, which is tedious and time consuming.

Developing a QA system to automatically provide brief and precise answer to a natural language question has been a long-standing research problem for its obvious practical and scientific value (Yih, Ma, 2016).An automated answering system is an important component of a distance teaching platform.

QA systems have been developed for a variety of purposes but still their level of accuracy, efficiency remains a problem and requires further research (Jilani, A., 2013).The focus was fundamentally around factoid questions, which constitute a small part of user information needs.One of the QA problems that have recently attracted a great deal of attention in the research community is the information source used for extracting answers (Yih, Ma, 2016.

The Quran is a vital book as it is a core text which contributes to the lives of millions of people today.There is a need for more research on Quran so as to make more sources available and accessible. Quran texts are interesting and challenging for evaluation researchers and Language resources. It can also have a great effect on society, helping the general public to access and understand religious texts (Atwell et al. 2012).). Question and answer corpus for the holy Quran is a valuable resource for the Question Answering research communities and may attract more useful research.

There is a need for a system to answer user's questions about the Quran.Our challenges are to specify exactly the requirements of knowledge-base, identify how to obtain this knowledge-base or make it available, and finally extract the desired answer that satisfies the user need.

1.3 Research Question/Hypothesis/Philosophy

1.3.1 Research Question

The research proposed will answer the following two questions. Firstly, is it possible to build a Question answering system based on advanced search techniques, tools,and specialized search dataset corpus? Secondly, how can these search techniquesand tools be used to build a system that returns an acceptable answer in an efficient and effective manner to a question posed by a user in order to fulfill his particular information need?

1.3.2 Research Hypothesis

Question Answering (QA) Systems that find an optimal answer to satisfy the user needs is a problem because mapping a user question to a piece of information (answer) -from a knowledge base- is difficult. Methods to find the optimal answer are needed. Knowledge base that is used in recent works has shown their shortage in solving the problem and finding the exact answer. More focus should be given to the knowledge base. Determining a good knowledge base can be used as resolution.

Obtaining a Closed-Domain Question answering system can be enhanced to give an accurate and direct answer by compilingspecific-domain knowledge base composed of questions along with their correct answers, and by applying sophisticated search approaches for both information retrieval and information extraction.

1.3.3 Research Philosophy

The Philosophy of this research is on automated Question Answering that tries to imitate the typical human Question Answering process of asking questions in natural language and receiving a correct answer. Asking questions can be done a plenty of different forms, systems that handle these different ways of asking questions can currently be built by collectingquestions from real world so the question can be found in different ways in differing contexts, which may optimize the performance of the Question Answering System.

1.4 Research Objectives

The main objective of this research is to investigate the techniques and tools necessary to build a novel QA system for the holy Quran using information retrieval, and natural language processing. These techniques and tools will aim to extract accurate answer to a question posed by a user. Additional objectives are listed as follows:

1. To understand the research problems by reviewing and evaluating existing QA systemsand computational research on Quran texts
2. Creating specific-domain knowledge base for the holy Quran by compiling bilingual dataset corpus composed of questions along with their correct answers from credible different sources.
3. Using data redundancy to improve the system performance.
4. To employ the best search techniques, tools andevaluate its accuracy.

The research outcomes achieved from the above objectives are:

- ✓ New data sources for the holy Quran which may be used by other researchers.
- ✓ A new application (question answering system)

Throughout this thesis, we showed how this aim and objective were fulfilled. A question answer pairs for the holy Quranwere collected, anda

QAEQ&AC (corpus dataset waswell prepared;to form a knowledge base for our system. This dataset is about1500 questions along with their correct answers. Data redundancy was used by reformulatingquestions. A question answering system for the holy Quran was designed tested and evaluated. The system performance was improved by applying data redundancy. The QAEQAS a Quranic Arabic/English Question Answering System was built using the python natural language toolkit (nltk) to process the user question and the corpus, as well as to implement the search engine to retrieve candidate results and then extract the best answer.

1.5 Research Methodology

The objectives of this study can be grouped into two main targets: (i) compiling the first Quranic Arabic/English Question and Answer Corpus(QAEQ&AC)and (ii) designing an automated Quranic Arabic/English Question Answering System (QAEQAS), which aims to map a user’s question to the most relevant question(s) from the Quran question and answer corpus, and then extract the answer for this question. We can drive the methodology that guided us to carry out this research in the following steps:

1. Investigation of the current research works that dealt with the area of QA system issues and challenges specially in Quran domain
2. Investigation of the techniques related to the area of QA system
3. Specification of the open issues that need to be addressed.
4. Identification and formulation of the research problem.
5. Development of the proposed solution’s architecture as follows:
    - Design the proposed solution.
    - Determine the appropriate methods and tools.

- Implementation of the proposed solution

6. Evaluate the overall solution

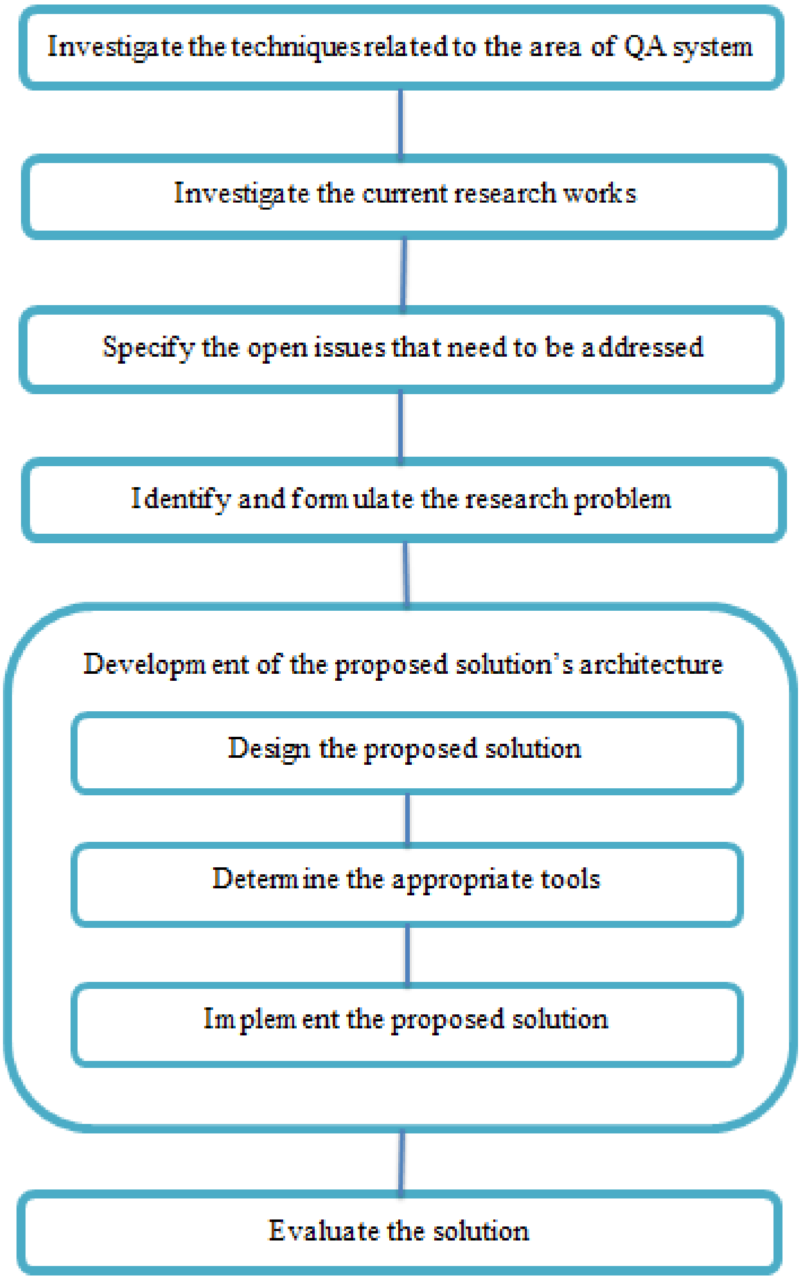

Investigate the techniques related to the area of QA system
Investigate the current research works
Specify the open issues that need to be addressed
Identify and formulate the research problem
Development of the proposed solution's architecture
Design the proposed solution
Determine the appropriate tools
Implement the proposed solution
Evaluate the solution

Figure 1.1Block diagram for research framework to carry the research objectives

1.6 ThesisContribution

To the best of my knowledge, there is no bilingual question and answer dataset corpus for the holy Quran has been compiled earlier. The work of this research lies in trying to apply existing advanced IR techniques coupled with NLP to a special domain of question and answer corpus using Python nltk to develop a question answering system for the holy Quran. This contribution can be summarized as follows:

- Compiling a bilingual Quranic Arabic/English Question and Answer Corpus(QAEQ&AC) datasetfrom scholarly sources: First the data were collected, and then prepared in many stages: merged, reformatted, cleaned, and then converted into the required format.
- Conducting machine learning experiments utilizing Quranic Q&A Corpus dataset to investigate their suitabilityin designing a Quranic question answering system or any other application where questions and answers are needed.
- Designing QAEQAS:a Quranic Arabic/ English Question Answering System.

Most of the research works presented in this thesis during the Ph.D have been published. Studies have been presented in the following Published papers:

- Quran question and answer corpus for data mining with WEKA. In 2016 Conference of Basic Sciences and Engineering Studies (SGCAC) (pp. 211-216). IEEE,IEEE Xplore digital library
- Using an Islamic Question and Answer Knowledge Base to answer questions about the holy Quran, International Journal on Islamic

Applications in Computer Science And Technology, Vol. 4, Issue 4, December 2016, 20 -29

- تجميع مدونة اسئلة و اجوبة للقرآن الكريم (Compiling a Quran Question and Answer Corpus)، الدورة العاشرة للمؤتمر الدولي لعلوم وهندسة الحاسوب (ايكا ICCA) بالتزامن مع الدورة الثالثة للمؤتمر الدولي لتقنيات المعلومات والاتصالات في التعليم والتدريب:(تِسات TICET) مارس 2016م (ICCA'2016 International Conference on Computing in Arabic, 2016)

- Evaluation corpus for restricted-domain question-answering systems for the holy Quran, International Journal of Science and Research (IJSR), Volume 6 Issue 8, August 2017

It is expected that contributions of this novel research will attract more researchers towards the subject, since we are planning to further extend the Quran question and answer corpus and make it freely available for reuse.

1.7 Research Scope

The work in this thesis covers the Quran as a whole, which makes it a massive domain for further work. The scope of this thesis is restricted to limited questions of the Frequently Asked Questions of the Quran. For example, detailed experimentation could be performed with the focus on 1500 question along with their correct answer. Most of these questions could be from community question answering (CQA) websites, which contain different types of questions and answers from real world, and cover the topics of interest within the Quran. It could be later enhanced to cover the questions about the Quran as a whole.

1.8 Thesis Outline

The rest of this thesis is organized as follows:

Chapter 2 is devoted to a background introduction of the main topics related to this thesis, particularly areas under data mining and processing. It also reviews literature on computing research on the Quran, and question answering system, investigates the existing solutions and gives more details about the research problem. Chapter 3 describes briefly the important parts of the methodology used in this thesis to achieve our contributions, and some theoretical underpinnings related to it. It includes an introduction clarifies the goals of our experiment and data gathering techniques. It also describes different concepts and methods of Information Retrieval and Natural Language Processing used by our methodology.

Chapter 4described in details how the proposed methodology has been implemented, the extensive methodology which includes all the pertinent information of the research: It gives a detailed description of the compilation of our knowledgebase source contributed through this thesis, namely the Quran Arabic/English Question and Answer Corpus (QAEQ&AC). It shows the automatic processing of Quran’s question and answers corpus using WEKA and NOOJ. Furthermore this chapter describes the steps used to create a Quranic Arabic/English Question Answering System (QAEQAS) using Python nltk in detailed. Chapter 5 presents discussion about the results and evaluation of the proposed solution. Chapter 6 concludes this thesis summarizing the main contribution and discussing potential work where this research could be extended in future.

1.9 Summary

The development of an automated question answering system depends on a good corpus; and various comprehensive technologies and tools. This research suggests a new QA system specialized for the Holy Quran. The system aimed for would accept a user question about the Quran, retrieve the most relevant questions, then extract the bestquestion from the Quran Questions corpus data set, and then retrieve the answer for that question.

This research focuses on building a special dataset resource that would eventually enter into beneficial text mining applications. The main contribution of this thesis has been the development of a novel resource, namely, a corpus of Quranic questions along with their correct answers compiled from scholarly sources. This resource is used in a question answering system for the holy Quran, and can be further used in text mining applications and machine learning experiments were Quranic questions and answers are needed.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

Question answering system can be defined as an automated system capable of processing a natural language question asked by a user, and retrieving an answer providing the requested information. Question answering tasks combines techniques from Artificial Intelligence (AI), Natural Language Processing (NLP), statistical analysis, pattern matching, Information Retrieval (IR), and information extraction (IE). This research is an inter-disciplinary project, proposes a methodology of Question Answering, gain benefit from existing advanced Information Retrieval and Natural Language Processing techniques using domain-specific resource to build a question answering system for the holy Quran.

Why question answering on the Quran

Quran is the holy book of Islam, it was verbally revealed through an angel Gabriel, to Prophet Mohammed (PBUH) before about 14 hundreds years, to be memorized by him and then passed on to others. This has been done gradually over a period of 23 years. Quran is written in Arabic language, it is the most truthful speech, it contains Allah’s message to all people. Quran is the main source of guidance and rules. It tells people how to act correctly and guide them to right way of life. The Holy Quran is the highest accepted source of religious legislation for around 1.5 billion Muslims around the world, Muslims read the Quran to understand the true teachings of Islam, where they find their religious teachings, knowledge and rulings. Muslims consider the Quran to be the only revealed book that has been protected by God from distortion or corruption.

This chapter serves as an introductory review, which describes relevant work: section 2 reviews some technologies of the main topics related to this thesis, used in QA systems, such as text mining, Natural language processing, pattern matching, information retrieval and information extraction. Anattention is also given to some text analysis tools like Knowledge Discovery (Data Mining) tools which semi-automate the process of discovering patterns in data. Section 3 discusses different approaches to the existing Quranic search techniques, text-based search techniques, and semantic search techniques which were used to search Quran and/or creating a QA system.

2.2 Background Studies and Related Technologies

2.2.1 Question Answering System

A Question Answering (QA) system is an automated approach to retrieve correct answers to questions written by users in natural language (Pujaret al, 2015). A QA system is a computer program that extracts its answers from a collection of structured (database) or unstructured (text) data known as the knowledge base. QA system enables users to access the knowledge resources in a natural way,by asking questions and to get back a relevant and appropriate response in a few words. The main goal of a QA system is to facilitate human-machine interactions, by getting quickly a precise answer rather than a list of documents.

To answer a question, a system must capable of analyze the natural language question, it has to find one or more answers by consulting existing knowledge base. QA systems are classified into two categories(Kangavari, et al, 2008): Closed domain and open domain QA systems. Closed-domain QA systems deal with questions in a specific domain such as Quran, medicine, or agriculture; closed-domain might point to a condition where only a limited type of questions are used, such as questions asking about description or fact. While open domain QA systems answer questions about anything in all domains and depend on general ontologies and world knowledge; moreover, open domain systems usually need much more data to extract the answer.

Even though QA systems have different applications, various design,but still there are some main concerns such as understanding a natural language question. In general, the processing of a QA system is composed of three stages: question analysis, document analysis, and answer analysis. The question is analyzed and classified to determine the question type and the major concept involved in the question; it is also reformulated to an appropriate query. The documents are analyzed to extract candidate documents that contain answers. The answer is analyzed to extract candidate answers and then rank to find the best one. Figure 2.1 shows the general Architecture of a Question Answering System.

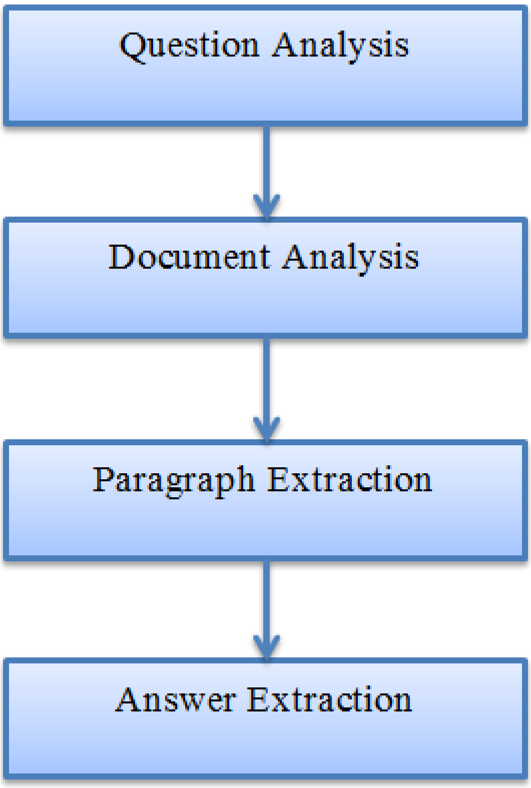


Figure 2.1General Architecture of a Question Answering System

The idealistic model of QA system should be able to understand a natural language question posed by a user, locate the desired information, extract this information from its sources and return it in the form of an answer. There has been decades of research aimed to achieve this goal and create this ideal QA system. Question answering (QA) received attention from the information retrieval, information extraction, machine learning, and natural language processing communities.

2.2.2 Natural Language Processing

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with interactions between computers and human natural languages. NLP is the computerized approach to analysing text that is based on both a set of theories and a set of technologies. Since text can contain information at many different level, from simple word or token-based representations, to rich hierarchical syntactic representations, to high-level logical representations across document collections. Developing aNLP application is a challenge because to the nature of the problems of complexity and ambiguity the natural Language poses.QA systems have recently become the focus of an increasing number of natural language processing projects. The computer normally requires humans to communicate to it in a precise, unambiguous and highly structured language.

Human speech, however, is not always precise, it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context (Rouse, 2017). Current approaches to NLP are based on machine learning, a type of artificial intelligence that examines and uses patterns in data to improve a program's own understanding. NLP is a technique that includes both natural language understanding and natural language generation. NLP is widely used to understand the meaning of the user query. Some of the most commonly tasks in NLP are: Automatic summarization, Machine translation, Morphological segmentation, Named Entity Recognition (NER), Natural language generation, Natural language understanding, Part-of-speech tagging, Parsing, Information retrieval, Information extraction, and Question answering.

2.2.3 Data Mining

Data mining is defined as the process of discovering useful patterns in data. The process should be automatic or semiautomatic (Witten et al 2011) Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, useful, and understandable patterns in data.Data mining is sometimes called knowledge discovery from databases. In knowledge discovery, what is retrieved is not explicit in the database rather, it is implicit patterns. Data mining finds these patterns and relationships using data analysis tools and techniques to build models. There are two main types of models in data mining. One is predictive models, which use data with known results to develop a model that can be used to explicitly predict values. Another is descriptive models, which describe patterns in existing data. All the models are abstract representations of reality, and can be guides to understanding business and suggest actions.

The two high-level primary goals of data mining, in practice, are prediction and description. The main tasks for data mining are:(1) classification: learning a function that maps (classifies) a data item (record or instance) into one of several predefined classes, (2) estimation: given some input data, coming up with a value for some unknown continuous variable, (3) prediction: same as classification and estimation except that the records are classified according to some future behavior or estimated future value, (4) association rules: determining which attribute or feature values typically go together in a data record or instance, also called dependency modeling, (5) clustering: segmenting a population of data records or instances into a number of subgroups or clusters, (6) Description and visualization: representing the data using visualization techniques, for human inspection of patterns.

Learning from data is categorized in two types: directed (supervised) and undirected (unsupervised) learning. Classification, estimation and prediction are examples of supervised learning tasks, while association rules, clustering, and description and visualization are examples of unsupervised learning tasks. In unsupervised learning, no variable is singled out as the target; the goal is to establish some relationship among all variables. Unsupervised learning attempts to find patterns without the use of a particular target field. Data mining steps in the knowledge discovery process are as follows (Jagtap, S. B., 2013):

3 Data cleaning: To remove noise and inconsistent data.
4 Data integration: To combine multiple sources of data.
5 Data selection: The retrieval of relevant data from the database.
6 Data transformation: The consolidation and transformation of data into forms suitable for mining.
7 Data mining: Using intelligent methods to extract patterns from data.
8 Pattern evaluation: To identify the interesting patterns

2.2.4 Text Mining

Data is unstructured text, and so text mining encompasses a new subfield of DM, focussing on Knowledge Discovery from unstructured text data.Most of text data faced on a daily basis is unstructured or free text, which is the more common type of text appearing in every day sources, unstructured text is the text that has not been processed into structured format (Miner, G., 2012).Text mining is known as intelligent text analysis, as text data mining or Knowledge-Discovery in Text (KDT) generally refer to the process of extracting useful and significant information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics.While structured data management is usually done using a database system, management of text data is done by the search engine due to the lack of structures. Search engines allow a user to obtain useful information from a collection conveniently with a keyword query (Aggarwal, et al., 2012)

Text mining is the process of deriving high-quality information from text, which derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text, (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modelling (i.e., learning relations between named entities).

Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods. A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted.

2.2.5 Machine Learning

Machine learning is the subfield of computer science, classed as a type of Artificial Intelligence. The basic task of machine learning is the construction of algorithms that can receive input data and use statistical analysis to predict output value within an acceptable range (Rouse, 2017). Machine learning provides computers with the ability to learn without being explicitly programmed, developed from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning finds the study and construction of algorithms that can learn from and make predictions on data, these algorithms overcome following robustly static program instructions by making data driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible.

Machine Learning allows computers to find patterns and predict classifications, clusters, associations and linking threads in a given set of data. The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, in data mining applications human uses his own understanding to discover pattern in data, while in machine learning patterns are detected automatically and the actions of the program are adjusted accordingly. Machine learning algorithms are frequently categorized as supervised or unsupervised. Supervised algorithms can apply what has been learned in the past to new data. Unsupervised algorithms can draw inferences from datasets.

2.2.6 Text-analysis Tools

2.2.6.1 Waikato Environment for Knowledge Analysis

The Waikato Environment for Knowledge Analysis (WEKA) is computer software developed at the University of Waikato in New Zealand. WEKA is free Java software available under the GNU General Public License. It is a collection of machine learning algorithms to solve data mining problems. WEKA has become one of the most widely used data mining systems while it offers many powerful features (Witten, Frank, 2005). WEKA support many different data mining tasks such as data preprocessing and visualization, attribute selection, classification, prediction, model evaluation, clustering, and association rule mining. It is written in Java and can be run with almost many computing platform. WEKA can be used to preprocess without writing any program code, and it comes with a graphical user interface to provide easily used tools for beginner users to identify hidden information from database and file systems in a simple way by using options and visual interfaces.

There is a specific default .ARFF data file format that WEKA accepts. The data should be as a single flat file or relation; the data can be imported from a Comma Separated Value (.CSV) file, a database, a URL etc.,where each data point is described by a fixed number of attributes. WEKA supports numeric, nominal, date and string attributes types. WEKA can be used to learn more about the data by applying a learning method to a dataset and analyze its output, and it is also used to generate predictions on new instances by using learned models, as well as to apply several different learners and compare their performance in order to choose one for prediction. The desired learning method is selected from a menu. A common evaluation module is used to measure the performance of all classifiers.

The most valuable resource that WEKA provides is the implementations of a wide range of data filtering tools, machine learning schemes, evaluation methods, and visualization tools. Filters are used to preprocess the data; we can select filters from a menu and then adjust their parameters according to our requirements. WEKA also includes implementations of algorithms for learning classifiers, association rules, clustering data for which no class value is specified, and selecting relevant attributes in the data. Also, there are many tools developed by third parties as add-ons. For example, WEKA was not designed for multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using WEKA. Another important area that is not covered by the algorithms included in WEKA is sequence modeling (Jagtap, 2013).

Clustering

Arranging data into reasonable groups is one of the essential methods of understanding and learning. Clustering is a technique in data mining where object are divided automatically into natural groups. These clustering algorithms attempt to identify similar objects and put them in one cluster. Clustering is an unsupervised Machine Learning technique where we only provide our dataset without class definitions, then the machine make natural divisions into two or more clusters. These clusters should then be inspected to find the relation between these objects (Muhammad A. B. 2012).

In general, clustering uses iterative techniques to group objects in a dataset into clusters that contain similar characteristics. These groupings are useful for exploring data, identifying abnormality in the data, and finally for making predictions. Clustering models also can help in identify relationships in a dataset that might not derived by browsing or simple observation. For these reasons, clustering is often used in the early phases of machine learning tasks, to explore the data.

The aim of clustering is to find structure in data and is therefore exploratory in nature. Clustering is widely used for a variety of machine learning tasks, such as detecting abnormal data, clustering of text, and analysis of a dataset prior to using in other classification or regression methods. In clustering the data that used to train the model were not need a label column. In other words, we don’t need to know any of the cluster categories in advance; the algorithm will find possible categories based just on the data.

Comma-Separated Values Files

Comma-Separated Values (.CSV) files are a common data exchange format that stores tabular data in plain text format (Hamoud and Atwell, 2016), which can be read using any standard text editor. CSVis supported by many applications and therefore a large amount of tabular data can be transferred between these applications. Each line of the CSV file is a record composes of one or more fields, separated by commas. Records are separated with end of line character, and this is used by the preprocess system. The tab-separated values and space-separated values are commonly used field delimiters in *CSV* files. *CSV* files are given the extension .csv. While there are various specifications and implementations for the CSV format, there is no official standard format and there are a variety of interpretations of CSV files. There is variation in the handling of fields which contain long strings, quote and double quote marks, and/or line-breaks; these are commonly found in Text Analytics datasets, where each data item may be a string representing a document, such as a news story, or a chapter or verse from a book. The format that is applied by most implementations is summarized as follows: (Dominic John Repici, 2010)

- Each record is placed on a separate line, delimited by a line break (CRLF**)**, but a record may span to more than one line when fields contain line-breaks (and field that containing line-breaks must be surrounded by double-quotes)
- The first record in the file may be a header record containing fields (columns) names, with the same format, and the same number of fields as the records in the rest of the file

- All records must contain the same number of fields throughout the file
- The last record in the file may or may not end with an ending line break
- Fields are separated with commas.
- The last field in the record must not be followed by a comma.
- Each field may or may not be enclosed in double quotes (some programs, such as Microsoft Excel, do not use double quotes). If fields are not enclosed with double quotes, then double quotes may not appear inside the fields
- A field which contains commas inside it must be enclosed in double-quote characters
- A field that containing line-breaks must be surrounded by double-quotes
- A field which contains double quote characters must be enclosed in double-quotes, and each of the embedded double-quotes must be also enclosed in double quotes

2.2.6.2 ARGO

ARGO is web-based text mining workbench, developed by The National Centre for Text Mining (NaCTeM) at The University of Manchester. It provides a graphical user interface through a web browser, which allows the user to create complex processing workflows composed of processing components and multiple branching and merging points. The Workflow processing is done on remote servers; the users can follow up the progress and see the results.

ARGO has a library of components comprising of natural language processing, text mining analytics, and user-interactive components such as

Manual Annotation Editor, which pauses the processing of workflows and waits for input from the user. Argo is a multi-user system, which shares workflows and documents between users and all eligible users can modify the shared objects. ARGO does not need installation procedures. However it needs an internet connection.

2.2.6.3 aConCorde

aConcorde is a free, multi-lingual concordance tool which was developed by AnderwRobers (2006) for native Arabic concordance. It has concordance functionality in addition to English and Arabic interfaces, is written in Java, supports Unicode (UTF-16), UTF8 and ASCII encoding, and is a frequency analysis tool, able to use different kinds of file types such as XML, HTML, RTF and Word files. It saves concordance output to a file (as plain text or HTML aligned tables), and saves the frequency list to file (as plain text or HTML table). However it ignores mark-up annotations within a corpus.

2.2.6.4 WordSmith

WordSmith (Scott, 2012) is a software package primarily meant for linguists, in particular for work in the field of corpus linguistics. It is a collection of modules aimed at searching for patterns in a language. It has a 'Keywords' function to compare corpora. It can be used for finding all instances of a word or phrase, and it helps find noticeable words in a text or set of texts, and then lists the words in alphabetical and frequency order. It also finds duplication of words, known as dependency words. The software is also available in several languages. Limitations include no real SGML/XML awareness, but there are possible solutions to this problem. Xml could for example be used first to annotate the text.

2.2.6.5 AntConc

AntConc (Laurence, 2011) is a freeware corpus analysis toolkit for concordance and text analysis developed by Laurence Anthony. It is fully Unicode compliant, works with all languages; allows full regular expressions for very complex searches, does word lists and keywords (by comparing against a reference corpus); does distribution plots of occurrences within each file; can handle lemma lists; and can handle XML-type and underscored tag-type part-of-speech tags. Furthermore the developer continually improves it and is open to feedback. Its disadvantages include it providing very minimal support for SGML/XML/HTML corpora, as it simply ignores rather than intelligently mines structural tags.

2.2.6.6 NooJ

NOOJ is a free linguistic software that created by Max Silberztein in 2002. NOOJ processes text and corpus to build concordances, it is also used as information extractor for search engines, text mining and intelligence applications. As a corpus processing tool, NooJ allows users to apply sophisticated linguistic queries to large corpora in order to build indices and concordances, annotate texts automatically, perform statistical analysis, etc. Linguistic modules can already be freely downloaded in many languages. Nooj characteristics:

- Can process texts in many different file formats, including HTML, PDF, MS-OFFICE, and all variants of UNICODE.
- Can import information from, and export its annotations back to, XML documents.

- Has an annotation system that allows all levels of grammars to be applied to texts without modifying them; this allows linguists to formalize various phenomena independently (Silberztein, 2008)

2.2.6.7 Gate Tool

GATE is free software under the GNU licenses and others, developed by The University of Sheffield (1995-2015), which can be used to solve a problem with text analysis and human language processing. GATE has been used for many IE projects in many languages and problem domains. GATE has a built-in IE component set called ANNIE, which is composed of many components included with GATE, such as Document Reset, Tokerniser, sentence splitter, RegEx Sentence Splitter, Part of Speech Tagger and semantic tagger.

2.2.7 Pattern Matching

Pattern matching is a mechanism for checking a value against a pattern. A successful match can also deconstruct a value into its constituent parts.In computer science, pattern matching is the action of checking a given sequence of tokens to find the constituents of some pattern, while in pattern recognition, the match usually has to be exact. Normally the patterns have two forms: sequences and tree structures. Uses of pattern matching include finding the locations (if any) of a pattern within a token sequence, to output some component of the matched pattern, and to substitute the matching pattern with some other token sequence (i.e., search and replace). Sequence patterns (e.g., a text string) are often described using regular expressions and matched using techniques such as backtracking. Tree patterns are used in some programming languages as a general tool to process data based on its structure.

Primitive pattern is the simplest pattern in pattern matching; it is an explicit value or a variable. More complex patterns can be built from the primitive patterns to form tree patterns. Pattern matching can be used to filter data of a certain structure. In symbolic programming languages, it is easy to have patterns as arguments to functions or as elements of data structures. A consequence of this is the ability to use patterns to declaratively make statements about pieces of data and to flexibly instruct functions how to operate. The most common form of pattern matching involves strings of characters. In many programming languages, a particular syntax of strings is used to represent regular expressions, which are patterns describing string characters.

2.2.8 Ontology

Ontology is a description of concepts with properties to be used in knowledge engineering as a knowledge base. Ontologies are used in information retrieval to retrieve more relevant information from a collection of unstructured information sources. Many ontology based information retrieval methods have been developed to make the information retrieval more effective, and these approaches can be categorized as: probabilistic, vector space, and semantic based.

2.2.9 Information Seeking

Information seeking is a form of problem solving (Marcus 1994, Marchionini 1992). It subjects the interaction among eight operations: problem understanding and admission, problem definition, selecting the search system, query formulation, query execution, checking the results, information extraction, and reflection/iteration/termination. In order to carry out effective searches the following user's experience have to be developed: knowledge about different sources of information, skills in defining search problems and applying search strategies, and efficiency in using electronic search tools. The information need, may be well or ambiguous defined by the user, so the analysis of the information need can be challenging, because users are faced with the general vocabulary problem as they are trying to translate their information need into a conceptual query. This is because a single word can have different meaning; as well the same concept can be described using different words.

Moreover, the concepts used for documents representation might be different from the user concepts. The conceptual query can take the shape of a natural language statement, so, it has to be translated to a new query that can be understood by the retrieval system. In a similar way, the documents have to be transformed in the form of text that can be processed by computer. The text can consist of multiple fields, such as the introduction, conclusion, and descriptor fields to catch the meaning of a document at different levels of resolution or focusing on different characteristic aspects of a document. When the user query has been executed by IR system, a user is offered with the retrieved documents. If these documents do not satisfy the user need, he modifies the query to start a new search. The modification of the query depends on user evaluation of the retrieved documents. Information retrieval is an interactive process.

In general information seeking involves data retrieval, or information retrieval, more specific information extraction. The widespread of huge amount of information on the internet, require the evolution of different systems and tools to simplify the retrieval, extraction, and searching for information or data from the internet and corpus. Information retrieval and information extraction are both associated with natural language processing, where computational linguistic techniques and theories play big role. IR is the process of retrieving relevant documents, which might be useful to user query. Information extraction process is a method for sorting a huge amount of text to extract relevant information and ignore the irrelevant ones. The IR system collects the relevant documents, and then the IE uses these documents to extract the relevant information

2.2.9.1 Information Retrieval

Since question answering is a special form of information retrieval, it’s reasonable to consider IR as an important technique sponsor in QA system .Information retrieval (IR) is the process of obtaining information resources relevant to an information need from a collection of information resources. An information retrieval process starts when a user posts his/her query into the system. A query does not uniquely identify a single object within the collection, but several objects with different degrees of relevancy may match the query. Information retrieval systems such as Google and Yahoo are used to remotely access and search a large information source, based on matching keywords and retrieving large amounts of information via web pages, with IR systems providing many matching algorithms, such as stemming query words, to rank the retrieved results.

Manning, C.D., Raghavan, P.and Schutze, H.,(2009) claim that: “Information retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)”. Information retrieval is quickly turned to the prevalent form of information access, overtaking conventional database style searching. The search can be relies on full-text or other content-based indexing. IR is also used to facilitate semi structured search such as finding a document where the title contains “python” and the body contains “matching". The information retrievalarea also covers supporting the browsing or filtering collections of document, furthermore processing a set of retrieved documents**.**

For efficient retrieving relevant documents by IR strategies, Users have to elicit their information need in a format that can be understood by the retrieval system, and the contents of large document collections need to be described in a form that allows the retrieval system to determine the likely relevant documents easily. Usually the user query is formulated, and the documents are transformed into a suitable representation, each strategy consolidates a certain model for its document representation purposes.

The evaluation of an information retrieval system is to evaluate how well a system meets the user's needs of information. Some evaluation metrics were designed for evaluating the performance of information retrieval systems, such as precision and recall as well more measures have also been proposed, generally, the measurement take into consideration a collection of documents to be searched and a search query, every document is known to be either relevant or non-relevant to a certain query. Modern information retrieval systems allow us to locate documents that might have the associated information, but the majority of them leave it to the user to extract the useful information from an ordered list.

2.2.9.2 Information Extraction

Information Extraction (IE) is to automatically extract structured information from unstructured and/or semistructured documents. More often this action interested in processing natural language texts by using natural language processing techniques. A main goal of IE is to do computation on unstructured data. In information retrieval automatic statistical methods have been developed to index large document collections and classify documents. The natural language processing is another complementary approach, which has solved the problem of modeling human language processing with large success. In terms of both difficulty and emphasis IE treats with tasks in between both IR and NLP. In terms of input, IE deals with a set of documents which have some similar entities or events, but the details are different, each of these documents has its own template.

The present significance of IE is related to the increasing amount of unstructured information. The Internet is considered as a web of documents, and more of the content is made available as a web of data. The web often consists of unstructured documents lacking semantic metadata, knowledge included within these documents can be made more accessible for machine processing through transformation into relational or XML forms. An ideal application of IE is to inspect a set of documents written in a natural language and establish a database from the extracted information. In order to apply information extraction on text, the text should be simplified by creating a structured view of the information existent in free text.

The overall goal is to create a readable text to be easily processed by machine. IE include: Named entity extraction: such as named entity recognition; Semi-structured information extraction: which tries to restore some kind information structure that has been lost through publication such as table extraction, Comments extraction; Language and vocabulary analysis: Terminology extraction such as finding the relevant terms for a given corpus. Many approaches combine multiple sub-tasks of IE in order to achieve a wider goal. Machine learning, statistical analysis and/or natural language processing are often used in IE

2.3 Quran Computing

Answering from the Quran is related to research on the Holy Quran, and question answering systems. Currently, many approaches to searching the Quran, and question answering systems exist. Some of these are discussed below:

The computational Quranic linguistics team at University of Leeds in the UK has accomplished a very distinct and notable work and techniques on Arabic language and Quranic field. Their most researches are about developing researching techniques and tools for the Holy Quran. Some of their current projects on the Quran are:Quranic Arabic corpus (Atwell et al., 2011), Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank,(Dukes, Atwell, Sharaf, 2010), QurAna: Corpus of the Quran annotated with Pronominal Anaphora(Sharaf, Atwell, 2012), QurSim: A corpus for evaluation of relatedness in short texts(Sharaf, Atwell, 2012), LAMP: a multimodal web platform for collaborative linguistic analysis (Dukes, Atwell, 2012), An Artificial Intelligence approach to Arabic and Islamic content on the internet(Atwell, et al, 2011), Arabic Language Computing applied to the Quran, and many more. They are currently working on a Quranic knowledge map, text mining hadith (Prophet Mohammad's teachings) and building Arabic question answering systems.

2.3.1 A New Question Answering System for Arabic Language

In the system(G. Kanaan et al. 2009) the user enters a short Arabic question; the system processes this question, and finds appropriate answers. The input for the QA system is a short question written in Arabic and a small set of ranked documents retrieved by the IR system. The System uses an existent NLP tagger to analyze both the users question and the documents, then it determines the assignment of each word to its root and its part of the speech, and finally it saves them in the project database. First the question analysis process is done by the NLP, then the system uses the IR system to retrieve candidate documents containing information relevant to the user's query; and after that the system chooses the most appropriate document according to the top similarity values, which are calculated by the IR system. The system uses keyword matching to generate the answer by matching the simple structures extracted from the question and the ones extracted from the candidate documents.

Their primarysource of knowledge is a collection of Arabic text documents. The researchers claimed that the overall success of the system was limited because the answer depended on the user query and the documents that contained the answer. Thus, if the user query was ambiguous; the retrieved documents would not contain the exact answer and would instead have to be ranked according to their relevance to the query. Sometimes users cannot find any answers for their questions, because the information needed is only retrieved from Arabic documents, meaning answers that may be found in other language documents are ignored. Furthermore the system does not cover all question types; rather it ignores some of them such as "how".

2.3.2 Semantic Query for Quran Documents Results

This system(M. A. Yunus et al. 2010) uses Quran documents in three languages (Malay, English, and Arabic): it translates words of an input query with semantics into another two languages, and then retrieves verses that contain words matching the translated words. It uses flat file documents and the searching process is done through pattern matching.

Formula is applied to test and evaluate query words and query numbers, and to calculate the percentage of precision and recall occurring. From their results it has been noticed that the system gives a better performance when retrieving more relevant document, compared to single query results. Unfortunately the system retrieval included some irrelevant documents which can go on to affect the system performance, and additionally it doesn't provide an actual question answering system just a way of ranking documents according to relevance to a inputted search query.

2.3.3 Quran ‘Search for a Concept’ Tool and Website

The creator(N H Abbas, 2009)developed an English/Arabic search tool for the Holy Quran which enhanced the recall to 87% and precision to 58%. ‘Mushaf Al Tajweed’ ontology of the Quran was used. The program covers about 1100 concepts in the Quran It extends the keyword search to synonyms and uses eight parallel English translations, with the search process also covering lemmas and morphemes to increase the search accuracy. The Quran corpus here consists of fifteen main concepts which cover all the themes of the Quran.

2.3.4 The Quranic Arabic Corpus System

In this system (Atwell et al., 2011**)** a Quranic online tool was designed that uses an Arabic language grammar based parsing for most of the main verses in the Holy Quran. It builds Quranic ontology of concepts based on the knowledge contained in traditional sources of Quranic analysis. However the system doesn't provide question answering tools.

2.3.5 Design of Automatic Question Answering System Base on CBR

This system (L Zhenqiu. 2012)would receive a question from a student, parse the question, extract the key words by thesaurus-based mechanical segmentations and then search the historical questions database according to the keywords to get the candidate questions. It would then calculate the similarity between the new questions and candidate history questions, and if the similarity reached the required threshold, the system would show the answers of retrieved history questions. If the difference between the questions and the history questions exceeded a certain threshold, the questions would be recorded in the answering database as a new question, and then end the program. If no similarity was achieved, the system would enter the full-text search module according to the keywords of the questions in order to search the full-text.

If the calculated similarity did achieve the required threshold, the system would display results of the full-text search; if not it would record the question and wait for answer to be provided. The researcher claimed that, the system is being continuously improved through the automatic adding of new questions and answers into the answering database, and it does not need a domain expert's interference. However the question used by this system should be short, and it being too long or complex may result in a wrong answer or no answer being provided at all.

2.3.6 Ontology Based Semantic Search in Holy Quran

In this system (H. Khanet al. 2013) a Quranic semantic search tool was suggested that worked by developing simple domain ontology for the Quran which was based on living creatures, including animals and birds that were mentioned in the Quran, in order to provide a Quranic semantic search. They built the ontology using the editor Protégé, and SPARQL was used to answer a query. The ontology provided 167 direct or indirect references of animals in the Holy Quran based on information found in the HewanatElQuran book. However their developed ontology is based on the domains animals and birds only, and does not cover all main domains in Quran, and it is not a functioning QA system.

2.3.7 AHybrid Approach for Question Classification in Persian Automatic Question Answering Systems

This system (E. Sherkatet al. 2014) combines rule-based and machine learning question classification approaches. Their approach is used practically in an online automatic question answering system named quranjooy. They created question taxonomy for the Quranic domain by collecting questions from the frequently asked question section (FAQ) of Quranic web sites. They defined a first version of their question taxonomy based on the knowledge of two Quranic expert and some existing question taxonomies, then manually associated these question taxonomy to the gathered questions. They then merged, deleted and amended some question classes and also added some new classes.

Finally they manually created their final version of taxonomy with their tagged classes by gathering more questions from credible websites and adding some new questions designed by expert, after which they associated the new questions created with the latest version of the taxonomy. Their taxonomy consists of 33 coarse grain and 75 fine grain classes. Their system architecture consists of: The Feature Extraction and Question Preprocessing; Rule-based Section; Machine Learning section. In The Feature Extraction and Question Preprocessing section the question being asked would be normalized and tokenized and then some feature would be extracted

In the Rule-based Section the question class would be determined, and questions that could not be categorized would be detected. In the Machine Learning section the question class would be predicted, and a feature vector would be prepared based on the features determined in the first section. The researchers claimed that their system used a high number of question classes, and they obtained high accuracy results. They also state that the machine learning section could be improved from time to time by considering the feedback of the users on the correctness or incorrectness of the detected class of the user's question. However, this would be very complicated work.

2.3.8 Semantically Answering Questions from the Holy Quran

This system (H. Shmeisani et al. 2014)finds the interpretation for user questions that are written in the Arabic language in order to find the best answer from the semantic representation of the Quran. They started by building an Arabic Ontology Extractor (AOE), which could create ontologies, from Quran Arabic texts. They detected about 380 concepts and 50 relations from Quran text. Examples of some concepts are: messenger, book, mountain and miracle. Examples of some relations are: live, create and show. To create instances of these concepts and relations they assumed that the relations are verbs, and the concepts are nouns. These two tags could then be detected from the POS-tagged Quranic text, and used to generate triples to store within the ontology.

The same patterns were also used in converting user queries. The developers specified three patterns that their system used to build the ontology, with other patterns to be added in the future. The patterns were based on the location of the verb with relevance to nouns within a sentence: a) two nouns followed by a verb, b) a verb followed by two nouns and, c) a verb between two nouns. The system was comprised of 3 layers: Question Preprocessing, Semantic, and Query Builder. In question preprocessing question cleaning is applied, and this involves finding the POS by using Standford parser, removing all stop words, and finding the answer type. After that the remaining concepts with their grammatical tags are passed to the semantic layer.

In the semantic layer synonyms for question words are created if they do not already exist in the ontology. Here the synonym dictionary at “AlMaany” was used to carry out the semantic feature for the system. In Query Builder POS tagging isused when building a SPARQL query which is run against the ontology to try and find an answer. The evaluation of their system is done by randomly choosing 35 asked questions. They claim that precision for this test data was 94%, and recall was 94%. However the current version of the system does not cover all relations and concepts found in Quran, moreover the system processes only factoid questions.

2.3.9 Al-Bayan: An Arabic Question Answering System for the Holy Quran

This system (H. Abdelnasseret al. 2014**)** prompts the user to enter an Arabic question about the Holy Quran. It retrieves relevant verses from the Quran along with their Arabic explanations from Tafseer books, then goes on to extract the text that contains the answer. Their corpus is made up of the Quran and its interpretation (two Tafseer books). The system integrated the Quranic Arabic Corpus Ontology (Dukes, 2013) and the Quranic Ontology (Abbas, 2009) to form 1,217 Quranic concepts, in addition to all the verses from the Holy Quran. The system has three phases: question analysis, IR, and answer extraction. In question analysis, the MADA toolkit is used (Morphological Analysis and Disambiguation for Arabic) (Habash et al. 2009) to produce part of a speech (POS) tag and a stem to each word included in the question.

Then all stopwords are removed so that they do not affect the information retrieval indexing like conjunctions, pronouns, prepositions, and other POS types. A support Vector Machine (SVM) is used to classify questions, though questions in which the question word is omitted can also be classified with SVM. In the IR stage, the system is based on the explicit semantic analysis approach (Gabrilovich and Markovitch, 2007), that improves keyword-based text representation with concept-based features, and where these features are automatically extracted from the Quranic Ontology.

Then the IR module retrieves related verses from the Holy Quran, and their interpretation from Tafseer books. In the answer extracted stage the named entities in the input question are identified, and then several features are extracted so that they can be used in the ranking of each candidate in order to extract the final answer to the input question. The researcher here claimed that the overall system accuracy reached 85%, using the top-3 results. However, it does contain some blemishes as some user queries may not match any concept from within the Quranic ontology used.

2.3.10 A Rule-Based Question Answering System on Relevant Documents of Indonesian Quran Translation

This system (RH Gusmita et al. 2014) was developed using a combination of two different architectures to complement each other: one of them used relevant documents and the other used a rule-based method. The rule based method used lexical and semantic heuristics to look for evidence that a sentence contained the answer to a question. For document analysis, the system implements a rule-based method on relevant documents, and adopts calculation to extract the correct answer.

The QA system produces question type, keywords, and keywords entity by analyzing user's question. The keyword is then used to find relevant documents from a related corpus, and then the retrieved documents are splatted into passages to easily find an answer. The named entity recognition would put a tag on every word contained in each passage that has the entity's name, and after that every passage would be scored based on the number of words that had a similar entity's name of answer type.

Finally, the system would find the answer from the passages that have a high-score. The researcher claimed that the system results indicated that it is not better than the previous one that used the rule-based method only. Furthermore this system is limited to processing only the first chapter in Quran (Al-Baqarah). Also, the rules were implemented for specific question types: (who, what, and where) and ignore other ones.

2.3.11 A Proposed Model for Quranic Arabic WordNet

The creators (M AlMaayah et al. 2014) proposed to develop a WordNet for Quran by building semantic connections between words in order to achieve a better understanding of the meanings of the Quranic words using traditional Arabic dictionaries and a Quran ontology. The Quran corpus will be used as text and Boundary Annotated Quran Corpus will be used to explore the root and Part-of-Speech for each word and the word by word English translations. Traditional Arabic dictionaries will be used to find the Arabic meaning and derived words for each root in the Quran Corpus. Then, these words and their meanings (Arabic, English) will be connected together through semantic relations. The achieved Quran WordNet will provide an integrated semantic Quran dictionary for the Arabic and English versions of the Quran.

2.3.12 Support Vector Machine Based Approach for Quranic Words Detection in Online Textual Content

The creator (T Sabbahet al. 2014) proposed a machine learning approach to detect Quranic words in a text extracted from online sources. They apply Support Vector Machine to generate a learning model of Quranic. Support Vector Machine (SVMs) is a set of supervised learning methods used for classification, regression and outliers' detection. (SVM) is based on the procedure of learning a linear hyper plane from a training set that separates positive examples from negative examples. The hyper-plane is located at the point in the hyperspace that maximizes the distance from the closest positive and negative examples (called support vectors).

Thus, SVM is designed to simultaneously minimize the empirical classification errors and maximize the geometric margin between positive and negative examples. There are several advantages of SVM as text classifier. First, SVM can handle exponential or even infinitely many features, because it does not have to represent examples in its transformed space, the only thing that needs to be computed efficiently is the similarity of two examples. Redundant features (that can be predicted from other features), and high dimensional features are well-handled, thus SVM does not need an aggressive feature selection.

Second, it processes error-estimating formulas, which can help SVM in predicting how well a classification is functioning. This eliminates the need for cross validation techniques. Filter and tokenizer were used to extract the textual content of the online resource. (Preprocess), then diacritics, and statistical features were performed. The JSP language is used for the user interface. Their system used a classification model in order to classify the words from online content. They used different features categories to develop prototype. Result gives high accuracy. However high complexity is involved in the process, more over the system doesn't provide question answering. Table 2.1 below summarizes some of these researches with their advantages and limitation.

Table 2.1 Summary of Quran computing systems

Index	**Paper title &Authors**	**Method**	**Advantage**	**Limitation**
1	A New Question Answering System for the Arabic Language (G. Kanaan et al. 2009)	Key word matching IR techniques joined with NLP approach.	Simple	Sometimes there is no answer or a wrong answer is generated. Does not cover all

				question types.
2	Semantic Query for Quran Documents Results (M. A. Yunus et al. 2010)	Semantic. Pattern matching search Cross-language information retrieval(CLIR) technique.	Totally relevant retrieval based on documents of three languages.	Retrieval of some irrelevant documents, the system doesn't provide question answering.
3	Design of Automatic Question Answering System Base on CBR (LIANG Zhenqiu. 2012)	Uses Case-based reasoning(CBR) techniques.	System intelligence is improved continuously. Does not need domain expert's interference.	May be gives wrong or no answer for long questions. Sometimes there is no answer for short questions.
4	Ontology Based Semantic Search in Holy Quran (H. Khanet al. 2013)	Semantic search. Ontology Based.	They created more classes to be expandable in the future.	Did not cover all main domains in Quran. Is not a QA

				system.
5	A hybrid approach for question classification in Persian automatic question answering systems(sherkat, E et al. 2014)	Combining rule-based and machine learning.	The machine learning section could be improved by considering the feedback of the users.	It is not a simple process, but rather very complicated and time consuming QA system.
6	Semantically Answering Questions from the Holy Quran (H. Shmeisani et al. 2014)	Semantic method.	Built the Arabic ontology extractor (AOE) to create Arabic ontology for the Holy Quran (reusable).	Does not cover all relations and concepts in the Holy Quran. Processes only factoid Questions.
7	Al-Bayan: An Arabic Question Answering System for the holy Quran. (H. Abdelnasseret al.	Semantic IR approach. Word matching. Machine Learning	Uses sophisticated tools such MADA for question reprocessing,	Some user queries may not match any concept from the Quranic

	2014)	techniques. SVM.	SVM for question classification.	ontology.
8	A Rule-Based Question Answering System on Relevant Documents of Indonesian Quran Translation (RH Gusmita et al. 2014)	Rule-based method.	Word match scoring function is used on relevant document only and not all documents, which can increase the systems performance.	Is implemented on only one chapter from the Quran (Al-Baqarah). Rules were implemented on who, what, and where, and neglected other question types.

2.4 Current Questioning Answer on Factoid Question

While answering all types of question is difficult, successful automated solutions to answer factoid questions have been recently developed by I.B.M. I.B.M has developed Watson Question Answering system to compete in Jeopardy-like games. The major approaches to answering factual questions range from shallow to deep. Shallow approaches use shallow linguistic methods combined with heuristic-based scoring techniques to locate and rank answers. Deep approaches rely on world knowledge and inference mechanisms to retrieve correct answers. There are several important outcomes resulting from these researches:

- The factoid questions and factoid Question Answering processes has become more understandable.
- Many useful Question Answering systems in different languages were developed.
- Search engines have become more advanced as a result of researches that have been developed. Now Search engines take queries in the form of natural language questions and provide a ranked list of short answers instead of just documents.

High-performance Question Answering systems have been built to be precise and fast enough to compete in real-time with top human contestants of the Jeopardy Game.

2.5 Conclusions: Quran Question Answering Systems

Question-Answering (QA) is an important research area, which is concerned with creating a system that automatically answers questions posed by humans in a natural language. By reviewing subsequent and more recent literature it has been found that the existing search techniques used in the Quran are Text-based search techniques or semantic search techniques. Text-based techniques can be a keyword-matching method which returns results that contain any query words. A Morphological-based method provides a root word search, where it generates all other forms of the query word and then finds all results matching these word forms.

The semantic search can be an ontology-based method that searches for the concepts matching a user query and then returns results related to these concepts. Cross-language information retrieval (CLIR) technique: translates words of an input query to another language, and then retrieves results that contain words matching the translated words. The synonyms-set method: produces all synonyms of the query word using WordNet and then finds results that contain words matching these words' synonyms or the query words.

Each of the existing techniques has their insufficiencies according to the retrieved results. The text-based search techniques retrieve some irrelevant results, while some relevant results are not retrieved. Furthermore the sequence of retrieved results sometimes are not in the right order, so the keyword-based technique's limitations include misunderstanding the exact meaning of the input words forming a query. Moreover, current semantic search techniques in the Quran have some limitations with regard to finding requested information. This is because these semantic searches use uncompleted Holy Quran ontologies. Additionally, these concepts have different scopes and formats.

Most of the recent works integrate two or more approaches to enhance their systems. It has been noticed that there are many differences in the content of the ontologies used by different systems, with some systems merging different translations of Quran ontologies, and others using only one language. The used ontology can cover the entire Quran in some systems while others use some parts of it, or only specific topics. Question class taxonomy may vary from one system to another. Some systems exploit general taxonomy for semantic classes like who, when, what, where and why type questions, while others utilize domain specific taxonomy. Also the tools used for building and representing the ontology can be different, and the evaluation methods that verify the ontology can be used in different ways (Alrehaili and Atwell, 2014).

A technical review of different QA system models and methodologies also shows that a typical QA system consists of different components to accept a natural language question from a user and return its answer back to the user. These components are used to understand the user question by applying a sequence of NLP processing operations, and then represent it in a format to be understandable by a machine, as well as can be mapped to an answer in the knowledge representation using different available techniques. In the recent research, researchers still agree that analyzing and understanding the meaning of a sentence written in natural language is by far one of the most difficult computing challenges.

To create QA systems a knowledge base is needed to extract the expected answer to the user question. This knowledge base is either domain-specific knowledge base for close domain systems that built specifically for the system, or domain-independent knowledge base for open domain systems such as the World Wide Web or a text repository. Different tools and techniques are available, but their selection is relied on the nature of the system under design.

From the previous sections in this chapter, many researchers have paid their attentions to answer questions about the holy Quran but none of them were use a corpus of questions along with their correct answer as knowledge base. The process of question answering on Quran's data can be considered as finding a domain specific special corpus to satisfy the query need. Our model is different in using specialized corpus composed of questions along with their correct answers as well as using data redundancy in order to improve the accuracy of the system.It extracts the best answer of the most similar question in the corpus as the answer to the new user question (Bogdanova, D., 2015)

CHAPTER III

PROPOSEDSOLUTION

3.1 Introduction

Developing a QA system to automatically provide brief and precise answers to a natural language question has been a long-standing research problem, for its obvious practical and scientific value. This research aims to create a QA system for the holy Quran that extracts accurate answer to a question posed by a user from its knowledge base. There are many issues that should be considered when designing a QA system such as the knowledge base: which provides a means for information to be searched. It is the source of knowledge likely to contain the answer to the question. Determine exactly the requirements of knowledge base and identify how to obtain this knowledge base. Furthermore consider the suitable tools and techniques that are used in data processing, searching, scoring and ranking by examining and exploring some alternative NLP technologies to investigate how suitable they are to build a QA system for answering questions about the Quran.

The main objective of this thesis is to develop a novel QA system for the holy Quran considering the knowledge base from which the correct answer for a question can be extracted, as well as the techniques used for searching and extracting this answer. In order to achieve this goal, knowledge base of questions along with their correct answers were compiled from a range of trusted sources. Then WEKA (Waikato Environment for Knowledge Analysis), NOOJ and Python NLTK tools have been examined and explored by conducting some experiments to investigate how suitable they are to build a QA system about the Quran. The rest of this chapter presents the proposed solution as well as describes briefly the methodology used in order to solve the research problem. Some theoretical underpinnings that are related to this chapter were also presented.

3.2 Theoretical Underpinnings

3.2.1 Knowledge Base

In general, a knowledge base is a central store for information: a public library, a database of related information about a particular subject that provides a means for information to be collected, organized, shared, searched and utilized. It can be designed either for machine or for human use. The knowledge base is the source of knowledge likely to contain the answer to the question. This knowledge can be structured in the form of a corpus such as the British National Corpus (BNC). It may be in the form of unstructured text such as on the World Wide Web or it can be constructed in machine understandable forms such as Databases and Semantic Web.

3.2.2 Corpus

“Corpus is a collection of writings, conversations, speeches, etc., that people use to study and describe a language “(Merriam-Webster). The corpus can be in written language or spoken language or both. Text corpus is a structured set of a large text stored in many files that share the same parameters, usually, the language, the structure and the encoding. Some popular corpora are British National Corpus (BNC), COBUILD/Birmingham Corpus, and IBM/Lancaster Spoken English Corpus. Corpus is used in many task such as hypothesis testing, statistical analysis, checking occurrences, and in the study of historical documents.

Monolingual corpora represent in only one language; bilingual corpora represent in two languages while multilingual corpora represent in multiple languages. To make the corpora more useful they are often undergoing to an annotation operation, such as morphological, Statistical and semantics analysis. Corpora are used as essential knowledge base in corpus linguistics. Corpus may be open or closed. The analysis and processing of different types of corpora are the theme of a lot of work in computational linguistics, machine translation and speech recognition such as part-of-speech tagging to signal the lemma form of the word.

3.2.3 Stop Words

Stop words are the most common words in any language such as short function words (“the”, “a”, “is”, “and”) in English language, which help build ideas, for example, linking sentences or words, but do not give any meaning. There are also some words within a domain specific corpus that may be appear very frequent, such words do not play a significant role in processing. Therefore these words are less considerableeither in the context of a language or in domain specific corpus. These words are considered as stop words, and need to be removed at some stage during the processing. During processing, the presence of these words may give misleading results by distorting the statistics in frequency-based methods as well consume valuable processing resources

There is no specific standard list of stop words that is used by all natural language processing tools. In fact not all tools use a stop words list. Some tools do not remove these stop words to help in phrase searching (Ullman et al. 2011). Any collection of words can be selected as the stop words for a specific purpose; it is not easy to determine the stop words, and on the other hand, stop words differ according to the case used. For example some search engines remove some of the most common words — including lexical words, such as "want"— from a query in order to improve performance. The English stop words list that included within NLTK can be imported to be used, as well as some extra Quran domain specific words can be added to this stop word. Arabic, stop words list are not included in NLTK; hence we can use any proper version of Arabic stop words list such as the one that is created by TahaZerrouki.

3.2.4 Tokenization

Tokenization is a method of breaking up a piece of text into many pieces, such as sentences and words. It is the process of splitting a string into a list of tokens. A token is a piece of a whole, so a paragraph is tokenized into sentences; a sentence is tokenized into words. In other words we can say a sentence is a token in a paragraph, and a word is a token in a sentence. This tokenized text or tokens are used as an input for further processing such as parsing or text mining. For example if we want to get to generalizations about words, such as the lemmas or parts of speech the first step we need to find the words. The tokenization occurs at the word level is sometimes difficult to define what is meant by a "word". Often a tokenizer relies on simple heuristics, for example:

- Punctuation and whitespace may or may not be included in the resulting list of tokens.
- All contiguous strings of alphabetic characters are part of one token; likewise with numbers
- Tokens are separated by whitespace characters, such as a space or line break, or by punctuation characters.

3.2.5 Similarity Measure

Similarity measure is a quantitative value that states the degree of similarity between two objects. In the case of a QA system the user question and the corpus document can be considered to be the two objects that need to obtain the degree of similarity between them. The value of similarity is between 0 and 1, when there is no match between the two objects the measure is equal to 0, which indicates irrelevant documents. The measure 1 represents the highest level of similarity, which might occurred when all the terms of user question are found in a document. The similarity measure is calculated for every document in the corpus, and then these documents are ranked using their relatedness to the user question. There are different methods to calculate the similarity measure, selecting a method depends on the nature of problem being solved.

3.2.6 Scoring

Candidates may have different importance in evaluating how a document matches a query. The candidates' results that obtainedafter searching can be classified according to the degree of match to the user query, a score can be given to each of these candidates according to the number of user query words it contains, and how close these words are to the candidate, the more and the closer the better. The score for candidate results can be calculated using different methods.

3.2.7 Ranking

A ranking is a relation between a set of items such that the items are organized in ascending or descending order according to specific criteria. Using rankings, information can be easily evaluated according to a particular norm, for example, the search engine ranks the pages according to their relevance to the user query, which makes it simpler for the user to select the pages they want. Documents candidates can be ranked based on numbers of keywords matched, measures of distance between keywords, and other similar heuristic metrics.

Ranking of query results is one of the essential problems in information retrieval. Given a query q and a collection D of documents that match the query, the problem is to rank that is sort the documents in D according to some criterion so that the "best" results appear early in the result list displayed to the user. Classically, ranking criteria are phrased in terms of relevance of documents with respect to an information need expressed in the query.

3.3 Proposed Solution

Most QA systems goal is to answer factoid questions, which identify some factual information, and since the questions in Quran domain, are normally searching for knowledge; our proposed methodology will answer all types of questions. This thesis proposes to use a corpus composed of questions along with their correct answers as knowledge base, as well as it benefits from data redundancy by reformulating the corpus questions in different ways in differing contexts to optimize the system performance. A major challenge in automated QA system that uses question and answer corpus, lies in identifying similar corpus questions to the user's questions. There is a need to find questions in a corpus that were semantically similar to a user's newly posed question. This will help in retrieving a high-quality answers that are associated with similar questions in the corpus as well as reducing the time lag associated with searching in CQA (community question answering) service (Li, S., 2011).

Our proposed solution uses a combination of advanced Information Retrieval and Natural Language processing tools and techniques to develop a Question-Answering system for the holy Quran. In this section we describe briefly the methodology used for the proposed solution. More discussion and details about this solution, and how this solution will be implemented to solve the research problem will be presented in chapter 4. Finallythe discussion about experimental results will be presented in chapter 5.

3.3.1 Preparing the Corpus

Most QA Systems mapping a user query to the most relevant document and identify the specific paragraphs or snips of the document that contain the information required to answer the question or contain the answer itself. While mapping a user's question to a piece of information from these documents is difficult, we used a corpus of question and answer which has been manually collected from a wide range of sources and designed to represent the Quranic Arabic-English Question and Answer Corpus (QAEQ&AC). It is a written bilingual corpus, which comprises Arabic and English text of question and answer pair. First, question-answer pairs have been collected from several trusted expert sources. Then a number of preprocessing steps were undertaken for the Quran questions and answers corpus: different data subsets were integrated by transforming multiple corpora into a merged form. After that the data were cleaned using Microsoft Excel to eliminate irrelevant and unwanted data, and then converted to a format that suitable for mining tools, where we have created a comma-separated value (CSV) file format.

3.3.2 Exploring WEKA and NOOJ

WEKA and Nooj have been tested and explored by conducting some experiment to investigate how suitable they are to build a QA system about the Quran. Some prototype experimental work have been done using

WEKA (Waikato Environment for Knowledge Analysis) to try and see if there are any problems which already exist, and to help me to identify technical issues. In doing this work, first some example of questions along with their answers have been selected from the Quran Question and answer corpus and then converted into a CSV (Comma Separated Value) format for a Machine Learning toolkit experiment. Finally this file format has been tested with WEKA; some clustering work has been done. Details have been provided in chapter 4 section 4.3.1.

We concluded that WEKA is good for classifiers and clusters but does not have functionality to build a question-answer interface. We then went on to investigate Nooj platform. A module forNooJhas been presented, which analyze, annotate and present the automatic processing of Quran question and answers corpus; details have been explained in chapter 4 section 4.3.2, again it is found that NOOJ is a good linguistic development environment that allows users to create formalized dictionaries and grammars and other theoretical work but it is not suited to a QA interface for Quran QA. So a third approach is to build my own implementation using Python and NLTK, this allowed me to develop a prototype QA system.

3.3.3 Developing QAEQAS Prototype

This thesis proposes a novel method utilizes the Quran Arabic-English Question and Answer Corpus (QAEQ&AC) as knowledge base and applying existing information retrieval and natural language processing techniques to develop a Quranic Arabic English Question Answering System (QAEQAS). QAEQASis unlike other question answering systems that focus on the generation of new answer; it is a natural language question-answering system that uses a corpus of question and answer as knowledge base to retrieve existing answers found in the corpus (Song W., et al, 2007)

QAEQAS aims to map a user query to the most relevant question from the Quran question and answer corpus, and then find the answer for this relevant question as answer for the user question. To achieve this goal a series of operations to the user question and documents (corpus questions) were applied to find ranked questions. As a complete question answering solution, python and toolkit were used to process the user question and the corpus, as well as to implement the search engine to retrieve candidates' results and then extract the best answer. These operations can be classified into fourphases. Each of them performs a specific operation to produce an output which is used as the input for the successive step. Figure 3.1 is the block diagram for QAEQAS.

1. Data Preprocessing
2. Information Retrieval
3. Scoring and Ranking
4. Answer Extraction

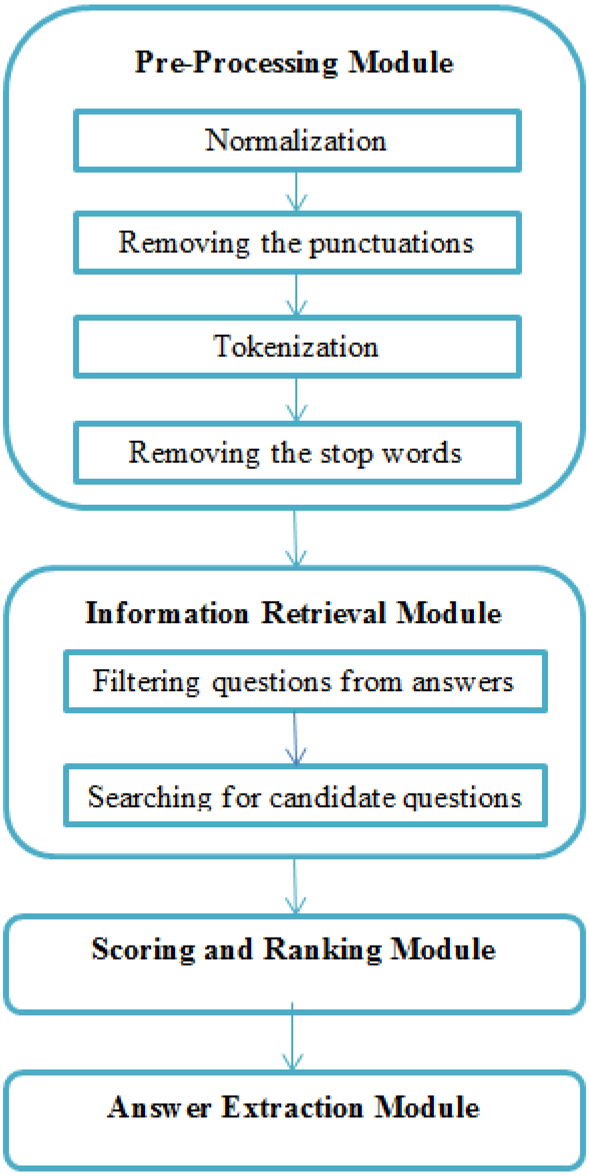

Pre-Processing Module
Normalization
Removing the punctuations
Tokenization
Removing the stop words
Information Retrieval Module
Filtering questions from answers
Searching for candidate questions
Scoring and Ranking Module
Answer Extraction Module

Figure 3.1 QAEQAS block diagram

1. Data Preprocessing

The first stepperformed in preprocessing is normalization: for English version the user question and the corpus questions were converted to lower case. For Arabic version normalization has been done in order to handle some problems with Arabic letters. Then the whitespaces and punctuation characterswere removed from user question and corpus questions for both Arabic and English versions. After that the user question wastokenized, by splitting it into linguistic units known as tokens. Then stop words wereremoved from the user question. Stop words are composedof existing stop words list that found in python nltk for English, and Arabic stop words list which created by TahaZerrouki, in addition tosome words in the corpus thatdeemed to be stop words. The remaining words are known as terms, which are used in searching the corpus questions.

2. Information Retrieval

In this phase the user question is mapped to corpus questions. This is done by comparing all terms of the user question, one by one, with every corpus question. This phase was used to retrieve all questions from the corpus questions that probably relevance to the user question, in other words all questions that contain one or more term of the user question were retrieved. These extracted questions are known as candidate questions. The objective of this phase is to identify if there is any document (question) that is relevant to the user question, and then include all relevant documents in a list of candidate question aggregation. As a result all irrelevant documents are discarded by filtering them out whereas all relevant question(s) are included in the candidate questions aggregation to be used for further processing to compete for more relevant question.

3. Scoring and Ranking

In this phase the similarity measure - the degree of relatedness of each candidate question to the user question- was obtained. In the LESK based similarity measure, raw document weight can be treated as a similarity measure. It is based on the argument that the higher the number of common terms between the user's query and a candidate result the higher the similarity measure. So a document having higher number of term is more related to the question than a document contains lower number of term, it depends on term count. In our method, document weighting depends on the number of terms in the document. A score was then given to each of these candidates according to the number of question terms it contains. Then we use a ranking method to identify the best matched document to a given user's question. So candidate documents are ranked according to their similarity measure. The higher the similarity measure, the higher will be the rank and vice versa. Candidate results are sorted in descending order of their similarity so that the best question is found at the top of the list.

4. Answer Extraction

In this final step ranked candidate documents determined in the previous step were used to pick up the top candidate question, which is the most related one to the user question. As we have explained before that the corpus file we used consists of lines, and every line is composed of question and answer. Our approach relies on finding the correct question from the corpus that matches the user question, and then the complete line which contains the best question is obtained. To obtain the answer Python's split() function have been used to divide the line and get the answer part, which represents an answer to the users question.

3.3.4 QAEQAS andData Redundancy

The knowledge base used by QAEQAS is relied on the data redundancy that already found in the corpus due to data collection from several different sources. In addition to that another set of questions were created to be semantically identical but syntactically different from the original questions, by manually reformulating each of the corpus questions into different contexts. In the first version of QAEQAS prototype, each question was reformulated so that each question has two extra variants in different context. In the second version of the prototype more data redundancy were used by reformulating each question into up to six different variants. This set was then added to the Quran question and answer corpus files; at the end our new data set files will list the original questions and the set of variants for that original along with their answers.

3.3.5 The Evaluation Method

Characteristics of QAEQAS have been investigated through user-oriented testing. Evaluating QAEQAS is based on domain expert knowledge, since the corpus is composed of pairs of questions, and its correct answers from credible sources. Since we do not have a Gold Standard for the Quran questions to compare with the results, we relied on an expert user group of Muslim university academics to ask the two versions of QAEQAS some questions in our questions domain. Details of evaluation method are presented in chapter 5.

3.4 Conclusions: Proposed Solution

This Chapter presents a brief summary ofthe research methodology for our proposed QA system, which based on Information retrieval, Natural language processing, and special corpus composed of question and answer pairs using the data redundancy, by adding differentvariants of the question. The chapter also describes different concepts and methods of Information Retrieval and Natural Language Processing used by our methodology. The next chapter presents more details about the proposed solution and how they will be implemented to solve the research problem.

CHAPTER IV

IMPLEMENTATION OF THE PROPOSED SOLUTION

4.1 Introduction

This chapter presents in details how the proposed methodology has been applied, as well as the idea behind the techniques tools and functions that have been used in order to carry out the proposed method. Furthermore, the chapter discusses the Python and natural language toolkit that has been used to implement the proposed methodology. Finally, the chapter presents a brief conclusion summarizing what have been done in this chapter.

The main objective of this thesis is to develop a QA system for the holy Quran, whichconsidering the knowledge base to return a correct answer and achieve better accuracy. This research, proposes tousea knowledgebase of question-answer pair and data redundancyaiming to minimize the retrieval of non-relevant document (questions) and hence retrieving the correct answer. This objective is achieved by applying a series of operations using information retrieval, natural language processing techniques and domain-specific resources. These operations can be classified into three distinct phases each of these phases performs a specific operations.

- Creation of an Islamic question and answer data set corpus.
- examination and exploration of some alternative NLP technologies:
    - ➢ Machine learning experiment using WEKA

- ➢ Using Nooj as an analysis tool for Quran question and answer corpus
- ➢ Parsing the Quran question and answer corpus using Python and Natural Language ToolKit (NLTK).

- Developing an automated question answering system for the holy Quran

Figure 4.1 is the block diagram which states the three main phases for the proposed solution

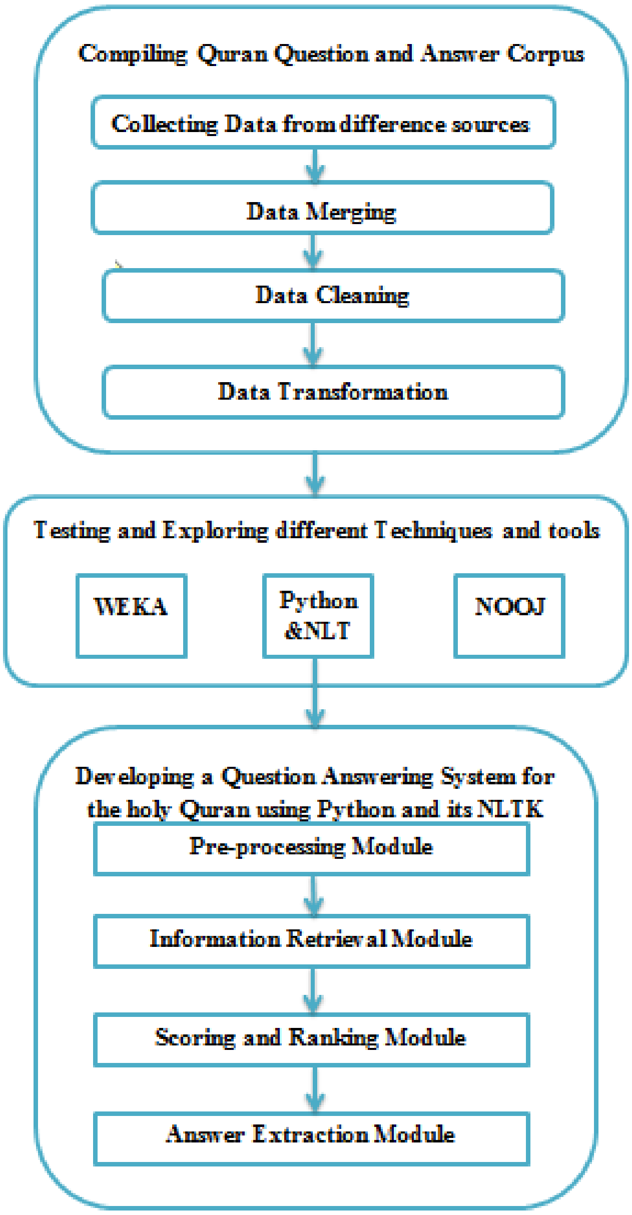

Compiling Quran Question and Answer Corpus
Collecting Data from difference sources
Data Merging
Data Cleaning
Data Transformation
Testing and Exploring different Techniques and tools
WEKA
Python &NLT
NOOJ
Developing a Question Answering System for the holy Quran using Python and its NLTK
Pre-processing Module
Information Retrieval Module
Scoring and Ranking Module
Answer Extraction Module

Figure 4.1 block diagram of the proposed solution

4.2 Creating an Islamic Question and Answer Data Set Corpus

4.2.1 Background

There is wider interest in evaluation of restricted-domain QA systems, in contrast to open-domain QA systems. A main characteristic of question answering in restricted domains is the integration of domain-specific information that is either developed for question answering or that has been developed for other purposes. We chose the Quran domain to develop a question and answer corpusfor the reason that :The Quran is the holy book of Islam, it contains Allah’s message to all humanity. The Quran is the main source of guidance and rules. Muslims read the Quran to understand the true teachings of Islam. There were no adequate existing resources specifically designed for Quran question and answer corpus. Existing Quran question and answer sources are scattered between different webpages, each of these sources has its own format and style. There is a need to create a unified dataset for Quran questions and answers, to be used in testing and evaluation in applications for Quran search and question-answering system.

In recent years, large amount of question and answer corpora have been searched and developed: De Smet, H.(2009) developed Yahoo-based Contrastive Corpus of Questions and Answers (YCCQA), it has been compiled using material downloaded from http://answers.yahoo.com/. The site offers an environment in which users post questions and answers. YCCQA contains about 90,000 question-answer pairs which make 29 million words of text. All the material collected has been posed by users between 2006 and 2009. Language material, consisting of questions and the accompanying answers, has been extracted for English, French, German, and Spanish. Ravichandran, D., Ittycheriah, A., &Roukos, S.

(2003) developed the KM database corpus, which is composed of question-answer pairs obtained from Knowledge Master (1999). Each of the pairs in KM represents a trivia question and its corresponding answer, such as the ones used in the trivia card game.

Soricut, R., & Brill, E. (2006) built a large training corpus consisting of question-answer pairs of a broad lexical coverage. They collected 1 million question-answer pairs. From FAQ pages this corpus is used to train various statistical models employed by their QA system in query analysis and answer extraction modules. Their system was intended to be applied to non-factoid questions. Burke, R. D., et al, (1997) developed a collection of approximately 30,000 question-answer pairs from the internet, obtained from more than 270 Frequently Asked Question (FAQ) files on various subjects. The obtained FAQ were used by their project FAQFinder.

Tomás, David, et al. (2009) developed a corpus of English question-answer pairs. The corpus consists of more than 70,000 samples. Each of these samples contains information that relates a question with its answer in four different contexts: exact match, sentence, paragraph and document. They claimed that their corpus suited to train on every stage of machine learning based QA system: question classification, information retrieval, answer extraction and answer validation. Feng et al., (2015) created InsuranceQA Corpus, a question answering corpus in insurance domain, contains questions and answers collected from the website Insurance Library. The content of this corpus consists of questions from real world users, the answers were composed by professionals with deep domain knowledge; this dataset is provided for research purpose only.

4.2.2 Community Question Answering

In order to create an accurate QA system, we need to integrate a wide range of knowledge, and use many ideas in natural language processing. Over the years of question answering research the focus was fundamentally around factoid questions, which constitute a small part of user information needs. Factoid questions are usually defined as questions that can be answered with a short phrase such as number. To solve the problem concerned with other types of questions beyond the scope of the factoid question is to ask community question answering (CQA) websites such as Islamic Knowledge, TurnToIslam,ALQuran, Islam web, etc. The CQA became widespread, and currently contain different types of questions and answers from real world. The majority of these questions can be classified as non-factoid questions (Jilani, A., 2013).

When more and more CQA services become available, the QA system can use question available, hence the CQA services can provide quite valuable resources for the QA system research(Li, S., 2011). Preliminary analysis of questions and potential answer sources gave an insight that the best data source is answers to similar questions in case they exist and we can find them. People often have similar tasks and situations which pose same questions. Therefore, it’s frequently the case that a similar question was already asked by someone and potentially even received a good reply and can be reused to answer new questions .As there are a lot of different types of questions that users post to CQA websites, our system is based on Frequently Asked Questions: a combination of CQA archive beside other sources.

4.2.3 Compilation of Quran’sQuestion and Answer Corpus

In the following sections we describe the compilation of the Quran Arabic-English Question and Answer Corpus (QAEQ&AC) through the integration of different data sets. QAEQ&AC is a written bilingual corpus which comprises question and answer pair in two languages, mainly Arabic and English. The logical basis behind this is to make the Quran question and answer corpus accessible to the English speakers, in the majority of Non-Arabic Muslim countries, such as Malaysia and Pakistan or any other people who speak English. Instead of applying sophisticated analysis our focus is on the data, so we used a specialized corpus of knowledge to find answers.

4.2.3.1 Data Collection and Sources

Data collection is the systematic approach to gathering information from a variety of sources to get a complete and accurate data of an area of interest. Accurate data collection is essential to maintaining the integrity of research, and to guarantee the quality assurance. The first step of building a corpus is to think about the resources of data. The process of collecting data can be relatively simple according to the type of tools used to collect the data. Data collection tools are used to collect information that can be used in many aspects such as evaluation of project performance. The collected data can also be reused for analysis purposes after refinement and cleaning. There are several tools that can be used to collect data.

The data collection tools should be good enough to collect useful data in order to have better evaluations for the research. Selecting specific tools depend on the nature of the task, as well as the type of the required data. Question-answering research in the religious domain involves ethical concerns. Answering questions about Islamic beliefs requires great care to give an accurate answer for a given question, which can be accepted religiously and universally. For instance, the answers should be evidenced as mentioned in Holy Quran as well as in Hadith books. Collecting questions and answers from authoritative and credible sources is an important issue; and low accuracies or wrong answers are not acceptable in the religious field especially in the Islamic domain.

Since there are not enough existing resources specifically designed for Quran questions and answers, we propose to merge different data subsets to comprise the Quran questions and answers dataset, as well as to have different questions from a range of different sources. Frequently asked question (FAQ) answering is a very useful in automated question answering systems, as it is effective to improve the efficiency of the whole system (Song W., et al, 2007). So,for our task we proposed to use Frequently Asked Questions (FAQs) from many sources to collect Quran questions and answers

In this thesis we selected fourtools to collect Quran questions and answers:

- A web-based tool, created by a group of scholars interested in the Islamic field.
- The Quran text book.
- Eliciting questions and answers from Muslims who came to the Holy Mosque in Makkah anda group of scholar from the Holy Mosque in Makkah
- From previous research.

Since the Internet is rich with data and easily accessible, the first and main source for our corpus collection is a web-based tool created by a group of scholars in the Islamic field (community question answering websites services). We proposed to focus on Frequently Asked Questions (FAQs), the reason behind this is that Muslims who use the Q & A services in these Islamic web sites to send their queries, give an insight into the practical aspects of the Muslim people who want to learn more about Islam, as well as non-Muslim people who want to know about Islam and compare it with different religions, maybe they want to convert to the religion that they deem appropriate from their perspective, and hence it can be argued that these are the topics of interest that need to be addressed. In order to collect the questions and answers from the web, we started by searching for web resources as raw data sources, the following are some of the selected web sites:

- TurnToIslam.com: is widely acknowledged to be one of the best places to learn about Islam, as it contains a huge library that covers many topics about Islam in many languages. It answers questions, shares Videos, Polls, Events and more. It aims at strengthening and uniting the nations and helping to show the beauty of Islam to the world, as well as building a kind, friendly community, on Islamic values.
- Islamic Knowledge/Come towards Islam which is another widely-used web site, which contains monthly archives covering many topics concerned with Islam such as questions and answers about the Quran, understanding Islam, Islamic facts, holy Quran chapters, teachings of the prophet Muhammad, haram (forbidden) food and drinks, Ramadan, women in Islam and many more topics. We examined its archives running from March 2011 till February 2015. This web site also provides the Quran text translated into many languages, as well as an Islamic resource for reading and listening to the Quran online with translation in various languages.

- All-Quran web site, which aims to have the holy Quran available to everyone in the world by having an easy way of audio streaming for a variety of Quran reciters and audio translations. It contains a tab for Islamic FAQ, in addition to further information about the holy Quran. It has been a commercial-free website since it was created.
- Siasat daily: The Siasat web site, which also provides questions and answers about the holy Quran. It is written in three languages: English, Urdu, and Hindi.
- Sultan Islamic linksDiscover Islam, Muslim People, Holy Quran and Islamic Religions. It has the linkIslam FAQ (Frequently Asked Questions) and the resource "you ask and the Quran answers"
- Islamic question and answer which contains many question about Islam written in 13 languages.
- A set of questions along with their answer were gathered from some forum such as the official forum for Sheikh Dr. Mohammad Al Arifi, and hussaballa forum.

Most of the content of the web based corpus consists of questions from real world users, the answers were given by professionals with deep domain knowledge.

We did not rely exclusively on web sources. A small set of questions and answers were gathered from some Muslims who came to the Holy Mosque in Makkah, and posed their questions to a group in the Holy Mosque who are leaders in the field of Islamic studies. Normally this group gives their expert answers to Islamic questions. As well as we included small test sets of questions and answers from previous research (Gusmita et al, 2014), (Abdelnasseret al, 2014), (Hamdelsayed& Atwell, 2016), (Shmeisania, H., et al, 2014) Furthermore some questions were taken directly along with their answers from the Quran text book, while others were extracted from it.

There are some questions along with their answers stated in Quran book, such as the questions that were directed to Prophet from different denominations - Muslims, Jews and Christians. These questions were taken directly along with their answers from the holy Quran book. For example : (يسألونك عن الأهلة قل هي مواقيت للناس والحج...) (يسألونك عن الخمر والميسر قل فيهما إثم كبير ومنافع للناس وإثمهما أكبر من نفعهما ...). وهنالك عدد من الأسئلة الاخرى مثل السؤال عن اصحاب الكهف وذي القرنين والروح...الخ such questions and their answers were subjected to reformulation, Table 4.1.shows some examples. Furthermore we extracted a set of questions and answers from some chapters of the Quran book, Table 4.2 shows some examples. To reformulate these questions and their answers optimally and add some explanation to it we used some interpretation reliable books such as Almokhtsar in the interpretation of the Quran, Tafsir Center for Quranic Studies. Then all these data sets were copied from their original sources and pasted into Microsoft word documents and then subjected to some preprocessing work.

Table 4.1 Examples of questions and answers directed to Prophet Mohammad

The location of the question and answer as stated in the Quran book	**Question after reformulation**	**Answer with some explanation from Tafsir books**
وَيَسْأَلُونَكَ عَنِ الرُّوحِ قُلِ الرُّوحُ مِنْ أَمْرِ رَبِّي وَمَا أُوتِيتُم مِّن الْعِلْمِ إِلاَّ قَلِيلا (الإسراء:85)	ما هي حقيقة الروح؟	لا يعلم حقيقة الروح إلا الله فهي سرٌّ من أسرار الله لم يطلع عليها عباده

وَيَسْأَلُونَكَ عَن ذِي الْقَرْنَيْنِ قُلْ سَأَتْلُو عَلَيْكُم مِّنْهُ ذِكْرًا إِنَّا مَكَّنَّا لَهُ فِي الأَرْضِ وَآتَيْنَاهُ مِن كُلِّ شَيْءٍ سَبَبًا (الكهف 83، 84)	من هو ذو القرنين؟	هو رجل طاف في الارض كلها وهو مثال للملك الصالح الذي آتاه الله ملكا فسخره في الدعوة لله تعالى وتعبيد الناس لهذا الدين
يَسْأَلُونَكَ عَنِ السَّاعَةِ أَيَّانَ مُرْسَاهَاۖ قُلْ إِنَّمَا عِلْمُهَا عِنْدَ رَبِّيۖ لَا يُجَلِّيهَا لِوَقْتِهَا إِلَّا هُوَۚ ثَقُلَتْ فِي السَّمَاوَاتِ وَالْأَرْضِۚ لَا تَأْتِيكُمْ إِلَّا بَغْتَةًۗ يَسْأَلُونَكَ كَأَنَّكَ حَفِيٌّ عَنْهَاۖ قُلْ إِنَّمَا عِلْمُهَا عِنْدَ اللَّهِ وَلَٰكِنَّ أَكْثَرَ النَّاسِ لَا يَعْلَمُونَ (الأعراف:187)	متى تقوم الساعة	علمها عند الله.وحده، لا يظهرها عند وقتها المقدر لها الا هو، خفى امر ظهورها على اهل السموات و اهل الارض، و لا تأتي الا فجأة

Table 4.2: Example, questions and answers extracted from Quran book

The location in the Quran book, where the question and answer were extracted	**Extracted Question**	**Extracted answer**
وَإِذْ قَالَ إِبْرَاهِيمُ لِأَبِيهِ آزَرَ ..... (الانعام:74)	ما اسم والد سيدنا إبراهيم عليه السلام	آزر
وَالسَّارِقُ وَالسَّارِقَةُ فَاقْطَعُوا أَيْدِيَهُمَا جَزَاءً بِمَا كَسَبَا نَكَالاً مِّنَ اللَّهِ وَاللَّهُ عَزِيزٌ حَكِيمٌ (المائدة:38)	ما هي عقوبة السارق والسارقة	يأمر الله تعالى بقطع يد السارق والسارقة
شَهْرُ رَمَضَانَ الَّذِيَ أنزِلَ فِيهِ الْقُرْآنُ..... (البقرة:185)	في اي شهر انزل القرآن الكريم	انزل القرآن في شهر رمضان

4.2.3.2 Data Preparation

Since there were no existing resources specifically designed for Quran Q&A, we merged different data subsets for the purpose of a Quran question and answer corpus task. Data preparation covers all tasks to build the final dataset from the initial raw data. Tasks include table, record, and attribute selection; merging; formatting; cleaning; and transformation for modeling tools. These tasks can be performed multiple times, and not in a specific order (Chapman, P., et al., 2000).

Creation of the Data File:

The obtained data must be processed or organized for analysis. For instance, these involve placing data into rows and columns in a table format for further analysis, such as within a spreadsheet. While our collected data sets have different format and style, a data file was created using Microsoft Excel 2010, and the collected data were transferred from Microsoft office word to Microsoft office Excel's documents by using the manual method "copy/Paste" and merged into standard spreadsheet worksheet format. After that their style and format were unified using Excel tools.

Data Cleaning

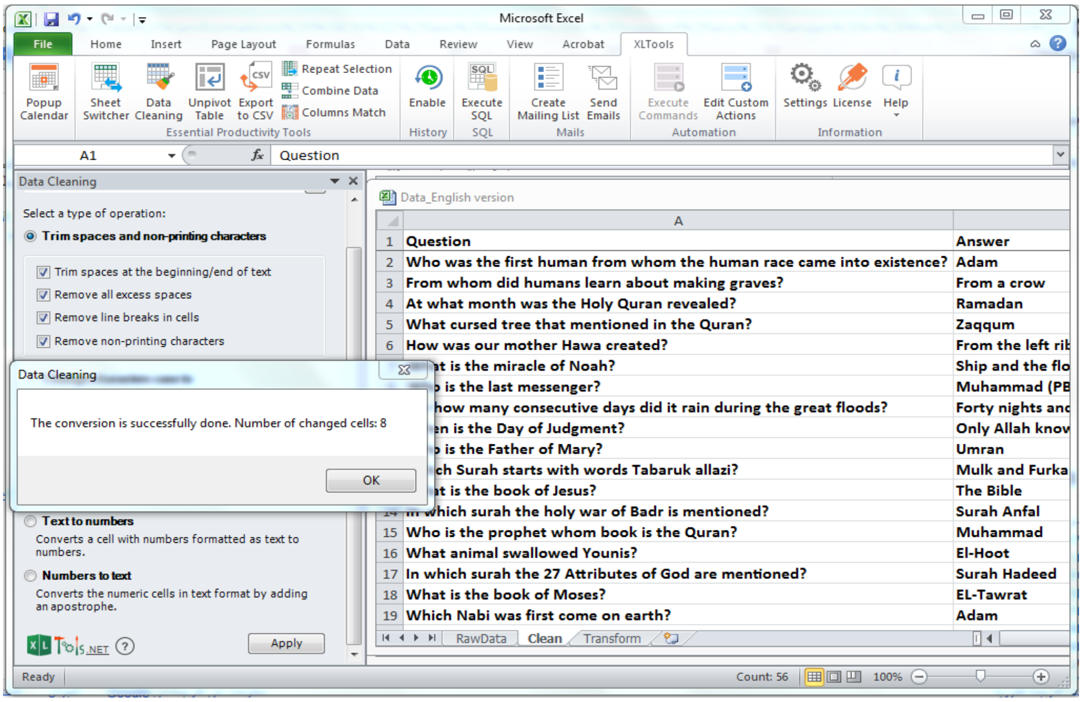
Data cleaning is the process of detecting and correcting corrupt or inaccurate data. Data quality problems exist in single data collections, such as files and databases, due to misspelled during data entry, missing information or invalid data. As for the data that is integrated from multiple sources, the need for data cleaning is largely increase due to the presence of heterogeneous data sources. Data cleaning is a phase in which noise and irrelevant data are removed from the collection; it refers to identifying incorrect, incomplete, inaccurate, etc. parts of the data and then replacing, modifying, or deleting this noisy data. Unclean data can contain mistakes such as punctuation errors, incorrect data associated with a field, incomplete or outdated data, or even data that has been duplicated in the database. Data cleaning may also involve activities like harmonization and standardization of data.

The goal of the data cleaning process is to maintain a meaningful data by removing elements that may hinder the analysis which affects the quality of the results, as the use of incorrect or inconsistent data may inevitably lead to false results. When we import or paste the data from the

Internet into Excel's worksheet, this may end up with unclean data, for example, may appear redundant spaces between words, and / or non-printable characters, etc., which will cause errors in the next phases. At this phase the original data were copied to another worksheet, and then cleaned by removing the inconsistent and the unwanted data from the dataset.

To clean data we used two methods: the manual method, and the automatic method. Manual method was used to remove incorrect or incomplete data while the automatic method may be not useful in such a situation. In automatic method XLTools was used to remove extra spaces, line breaks, delete non-printable characters etc., while the manual method may take painstaking hours or may not guarantee to detect and remove all the mistakes. We also handled the blank cells as they can create problems if not treated in advance. The duplicated data were removed as well as spell checker were applied on the data set to correct spelling errors, where it is nothing that reduces the credibility of the work more than spelling error.

While the search and replace in data cleaning is indispensable, we searched for inappropriate words and then replace them with more proportional ones. The text can also be changed to the uppercase, lowercase or other common capitalization, as well as the transformation of cell formats can be applied to change numbers to text and vice versa. Figure 4.2 bellow shows data cleaning using Excel tools (xlTools). Figure 4.3 and Figure 4.4 show examples of Arabic and English data respectively after cleaning.



A1 Question

Data Cleaning

Select a type of operation:

● Trim spaces and non-printing characters

☑ Trim spaces at the beginning/end of text

☑ Remove all excess spaces

☑ Remove line breaks in cells

☑ Remove non-printing characters

○ Text to numbers

Converts a cell with numbers formatted as text to numbers.

○ Numbers to text

Converts the numeric cells in text format by adding an apostrophe.

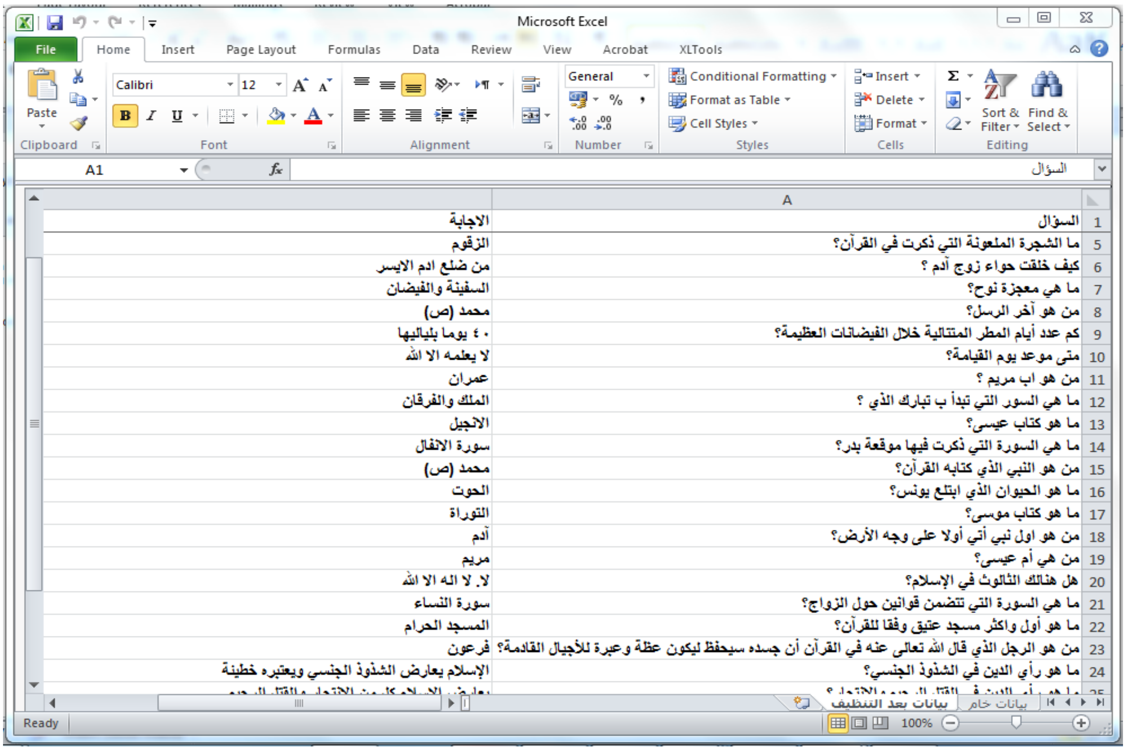
Data Cleaning

The conversion is successfully done. Number of changed cells: 8

Data_English version

	A	
1	Question	Answer
2	Who was the first human from whom the human race came into existence?	Adam
3	From whom did humans learn about making graves?	From a crow
4	At what month was the Holy Quran revealed?	Ramadan
5	What cursed tree that mentioned in the Quran?	Zaqqum
6	How was our mother Hawa created?	From the left rib
	at is the miracle of Noah?	Ship and the flo
	is the last messenger?	Muhammad (PB
	how many consecutive days did it rain during the great floods?	Forty nights and
	en is the Day of Judgment?	Only Allah know
	is the Father of Mary?	Umran
	ch Surah starts with words Tabaruk allazi?	Mulk and Furka
	at is the book of Jesus?	The Bible
14	In which surah the holy war of Badr is mentioned?	Surah Anfal
15	Who is the prophet whom book is the Quran?	Muhammad
16	What animal swallowed Younis?	El-Hoot
17	In which surah the 27 Attributes of God are mentioned?	Surah Hadeed
18	What is the book of Moses?	EL-Tawrat
19	Which Nabi was first come on earth?	Adam

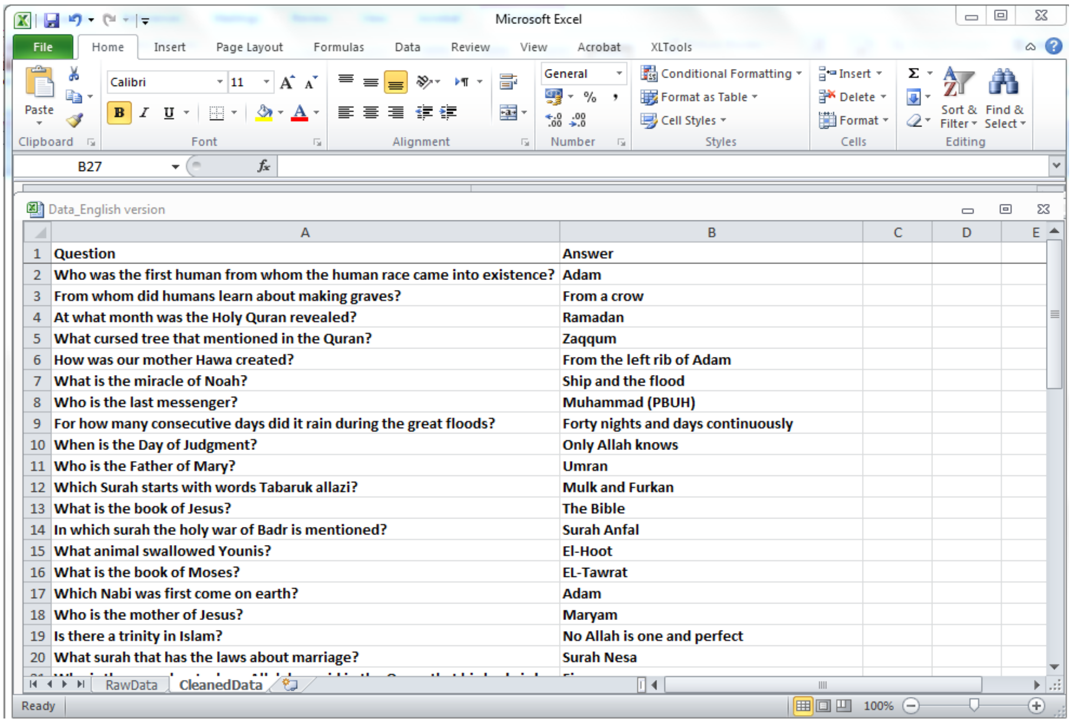
Figure 4.2data cleaning using Excel tools (xlTools)



A1 السؤال

	A	
1	السؤال	الاجابة
5	ما الشجرة الملعونة التي ذكرت في القرآن؟	الزقوم
6	كيف خلقت حواء زوج آدم ؟	من ضلع ادم الايسر
7	ما هي معجزة نوح؟	السفينة والفيضان
8	من هو آخر الرسل؟	محمد (ص)
9	كم عدد أيام المطر المتتالية خلال الفيضانات العظيمة؟	٤٠ يوما بلياليها
10	متى موعد يوم القيامة؟	لا يعلمه الا الله
11	من هو اب مريم ؟	عمران
12	ما هي السور التي تبدأ ب تبارك الذي ؟	الملك والفرقان
13	ما هو كتاب عيسى؟	الانجيل
14	ما هي السورة التي ذكرت فيها موقعة بدر؟	سورة الانفال
15	من هو النبي الذي كتابه القرآن؟	محمد (ص)
16	ما هو الحيوان الذي ابتلع يونس؟	الحوت
17	ما هو كتاب موسى؟	التوراة
18	من هو اول نبي أتي أولا على وجه الأرض؟	آدم
19	من هي أم عيسى؟	مريم
20	هل هنالك الثالوث في الإسلام؟	لا. لا اله الا الله
21	ما هي السورة التي تتضمن قوانين حول الزواج؟	سورة النساء
22	ما هو أول واكثر مسجد عتيق وفقا للقرآن؟	المسجد الحرام
23	من هو الرجل الذي قال الله تعالى عنه في القرآن أن جسده سيحفظ ليكون عظة وعبرة للأجيال القادمة؟	فرعون
24	ما هو رأي الدين في الشذوذ الجنسي؟	الإسلام يعارض الشذوذ الجنسي ويعتبره خطيئة

Figure 4.3 examples of Arabic data after cleaning



	A	B	C	D	E
1	Question	Answer			
2	Who was the first human from whom the human race came into existence?	Adam			
3	From whom did humans learn about making graves?	From a crow			
4	At what month was the Holy Quran revealed?	Ramadan			
5	What cursed tree that mentioned in the Quran?	Zaqqum			
6	How was our mother Hawa created?	From the left rib of Adam			
7	What is the miracle of Noah?	Ship and the flood			
8	Who is the last messenger?	Muhammad (PBUH)			
9	For how many consecutive days did it rain during the great floods?	Forty nights and days continuously			
10	When is the Day of Judgment?	Only Allah knows			
11	Who is the Father of Mary?	Umran			
12	Which Surah starts with words Tabaruk allazi?	Mulk and Furkan			
13	What is the book of Jesus?	The Bible			
14	In which surah the holy war of Badr is mentioned?	Surah Anfal			
15	What animal swallowed Younis?	El-Hoot			
16	What is the book of Moses?	EL-Tawrat			
17	Which Nabi was first come on earth?	Adam			
18	Who is the mother of Jesus?	Maryam			
19	Is there a trinity in Islam?	No Allah is one and perfect			
20	What surah that has the laws about marriage?	Surah Nesa			

Figure 4.4 examples of English data after cleaning

Data Transformation

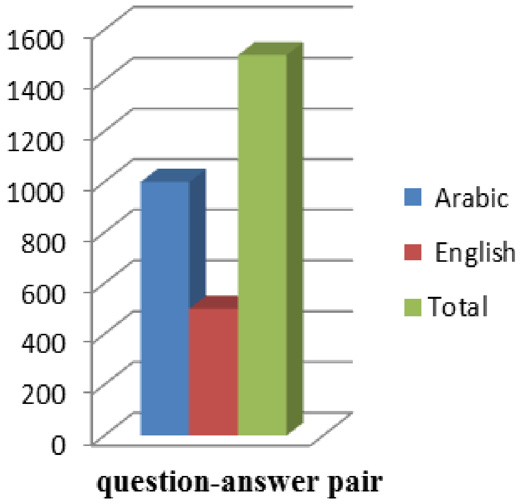
Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system. Data transformation allows the mapping of the data from its given format into the format expected by the modeling tool. Excel is a great tool to use when we need to take the data in a specific format, then processed to be converted into another format, then push the results to another tool to further processing. For example if we want to export Excel file to other applications that support the comma-separated value format (CSV) format, we can convert the worksheet first to CSV format and then export the .csv file to those programs. Excel works as a transfer tool for the data to be transferred from one system to another, where it supports many file formats.

In this phase the cleaned data were copied to another excel workbook, because CSV format does not support workbook containing several worksheets. After that the workbook was saved as a CSV format. CSV files are a common data exchange format that stores tabular data in plain text format, which can be read using any standard text editor. CSV is supported by many applications and therefore a large amount of tabular data can be transferred between these applications, (Hamoud and Atwell, 2016) whichincreases its popularity and its ability to survive, at least as an alternative format to import and export data. Since the CSV files are plain text, this makes it easy to be understand by normal user or even by beginner, as well as it allow users to diagnose data problems easily.

4.2.4 Result: QAEQ&AC

A Quranic Arabic/English Question & Answer Corpus (QAEQ&AC) has been developed. QAEQ&AC is a written bilingual corpus, comprises Arabic and English text. In its current form, QAEQ&AC contains about 1500 question-answer pairs, which makes about 42500 words of text. These questions and answers are divided over the Arabic and English languages as follows: 1000 Arabic question-answer pairs, which make about 20.000 words, and 500 English question-answer pairs, which are about 22500 word, Figure 4.5 and Figure 4.6 show the questions and answers, and the word count by each language respectively. Contrary to other contrastive corpora, this version of our corpus does not contain parallel translated texts. All texts are originals. As a result, the subcomponents of the corpus can be used independently as language-specific corpus.

The QAEQ&AC is not a general corpus; it is specifically dedicated in Quran domain, it includes different question types such as what, when, why, etc., Table 4.3 shows different types of questions. The answers can be in different length, a short answer, or longer text for questions that need more explanation. We anticipate that the current and subsequent versions of our corpus will be a valuable evaluation resource for computational linguists investigating Quran question and answer; it might be used as a gold standard in researches, that dealing with natural language processing, information retrieval, artificial intelligence. This corpus can be subjected to an annotation to derive linguistic information such as morphological, syntactic, semantic, and lexical information.



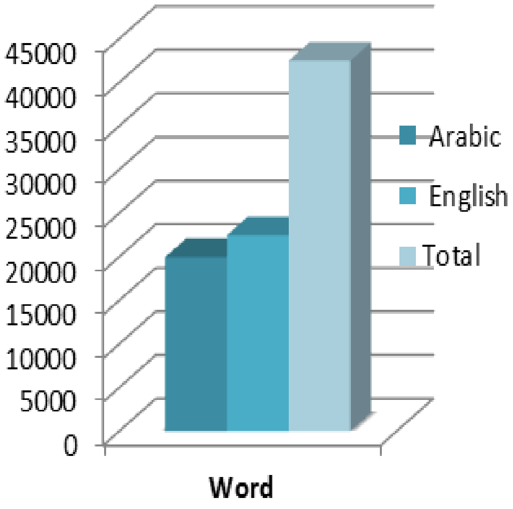


Figure 4.5questions and answers by language

Figure 4.6 word count by language

Table 4.3 question types

Domain Identifier	**The Meaning**
What	Indicates a question asking about things
Who	Indicates a question asking about persons
When	Indicates a question asking about time

Where	Indicates a question asking about a place.
How many	Indicates a question asking about countable things
Why	Indicates a question asking about cause
How	Indicates a question asking about the condition or situation
Others	All other questions

4.2.5 Preparation for further Investigation and Analysis

Text analysis is a set of linguistic, statistical and machine learning techniques that contribute to the extraction of informational content in text sources. Text analysis indicates a derivation of high-quality information from text, and this is done through devising patterns. Text analysis involves the retrieval of information; lexical analysis to study the frequency distribution of words; pattern recognition; tags/annotation; information extraction; and data mining techniques, whichinclude visualization, and predictive analysis. The main objective of these operations is to convert the text into data that can be analyzed, through the application of natural language processing (NLP) and analytical methods. This information content is used in many areas such as research and studies or exploratory analysis of the data.

Typical tasks of text mining include classification and aggregation, extract concepts, etc. Some preprocessing steps can be applied for the data to be ready for further analysis. For example WEKA filters can be used for preprocessing the data such as removing a certain attribute or removing instances that meet a certain condition, or to convert the attributes from a type that WEKA fails to tokenize and mine to another type (the correct type) that it can work on by using filters such asStringToWordVector.

After data preprocessing data mining techniques such as classification, clustering, visualization can be applied.

4.2.6 Conclusions: QAEQ&AC

In this section, we described the compilation of the Quran question-answer pair's collection, through harvesting data from websites, Islamic experts, and existing research datasets. The corpus has been manually collected from a wide range of sources, and designed to represent the Quran Arabic-English Question and Answer Corpus (QAEQ&AC). QAEQ&AC is a written, bilingual corpus, which comprises Arabic and English text. First, question-answer pairs have been collected from several trusted expert sources. Then the data were merged and cleaned using Microsoft Excel. After that data were converted to the format that suitable for mining tools, where we have created a comma-separated value (CSV) file format.

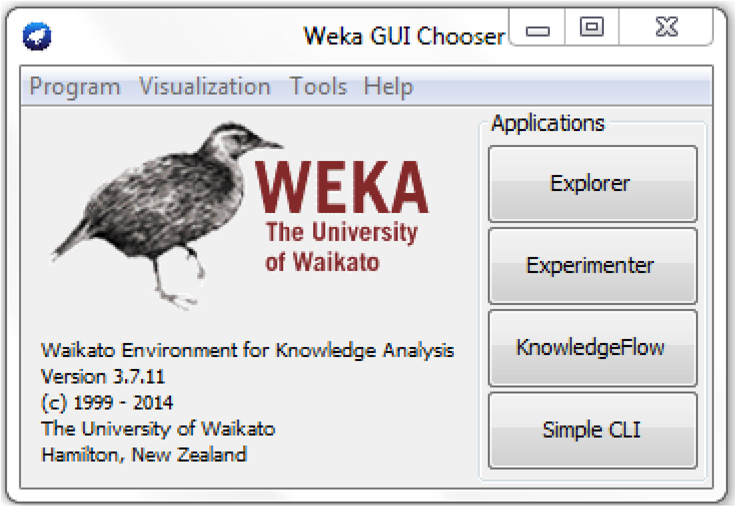
The corpus obtained consists of about 1500 question-answer pairs, which is equal42.500 words, divided over Arabic and English languages. It includes different question types such as what, when, why, etc., and different answer length. Collecting data manually is a big challenge, as the automatic method has not always been successful in filtering inappropriate or unwanted data. We believe that creating an integrated Quran question and answer corpus dataset is an important resource that we would like to apply in a task challenge, aimed at improving the state-of-the-art of online Quran question answering systems.

4.3 Using Text Analysis Tools to Analyze Quran's QA Corpus

After the Quran's Q&A corpus were compiled, our second step is to investigate some advanced computerized automated analysis tools for automatic processing of our written corpora and to develop a question answering system for the holy Quran. In this research we investigated three text analysis tools mainly WEKA, NOOJ and Python NLTK to analyze the Quran's QA corpus.

4.3.1 Machine Learning Experiments using WEKA

WEKA (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms to solve data mining problems. The WEKA Graphical user Interface (GUI) Chooser provides a starting point for launching WEKA's applications, this GUI Chooser consists of four buttons, to start applications (Remco R., et al ,2014). Figure 4.7 shows The WEKA Graphical user Interface Chooser. The main application is the Explorer, which explores data with WEKA. The simple command line interface (SimpleCLI) allows direct execution of WEKA commands for operating systems that do not provide their own command line interface. KnowledgeFlow supports the same functions as the Explorer but with a drag-and-drop interface, and it provides a framework for incremental experiments in machine learning. The Experimenter is used to carry out experiments and perform statistical tests between several learning schemes.



Waikato Environment for Knowledge Analysis
Version 3.7.11
(c) 1999 - 2014
The University of Waikato
Hamilton, New Zealand

Figure 4.7the **WEKA** Graphical user Interface

The Explorer is the most used tool, and is composed of several panels to allow access to the main components of the workbench : (1) the Preprocess panel which is used to import data , and preprocess this data by using filters to transform the data and prepare it according to specific criteria, (2). the Classify panel which allow applying classification and regression algorithms to a dataset, (3) the Cluster panel enables access to the clustering techniques in WEKA, (4) the Associate panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data, (5) the Select Attributes panel gives algorithms for identifying the most predictive attributes in a dataset, (6) the Visualize panel shows picture representations of data and results, such as a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

The first four buttons at the top of the Preprocess section enable us to load data into WEKA, by importing it from any file in the default ARFF format supported by WEKA, or any other format accepted by WEKA for which a filter is implemented, such as Excel Comma Separated Value (.CSV) file, a SQL database, a URL, etc., for preprocessing.

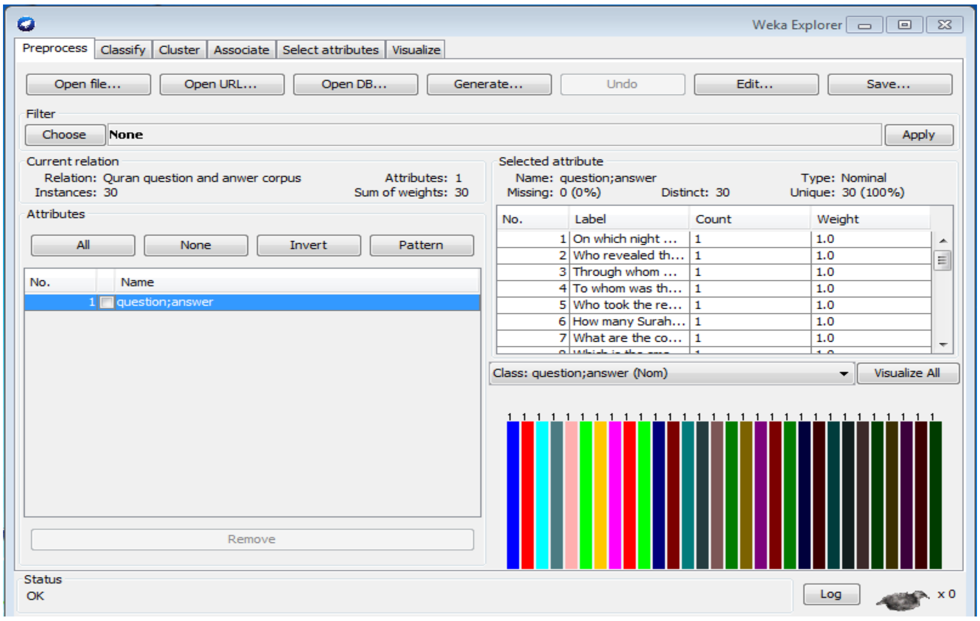
4.3.1.1 Preparing CSV File for WEKA

A representative sample of 30 questions along with their answerswere selected from the collected questions and answers dataset, to include representative samples covering all methods and sources that were used to collect them, and the questions types, as well as the length of the answer. There are some questions that need a long explanation for their answers. The selected dataset required cleaning prior to data usage; therefore the data were cleaned according to WEKA needs by removing control characters, and resolving formatting problems concerned with some characters such as double quotes, single quotes, comma, apostrophe, etc. .this sample were entered in an Excel worksheet and then converted to a CSV file

4.3.1.2 Loading the Corpus into WEKA

To load the CSV file, from WEKA chooser GUI we selected Explorer application button, and then the Preprocess panel, which is used to choose and modify the data being acted on. After that we chose the Open File button to display a dialog box allowing browsing for our CSV data file. From the dialog box we selected Open button to load our file into WEKA. WEKA also enables us to load data from other locations by selecting the desired button. The Open URL button allows asking for a Uniform Resource Locator web-address where the data are stored. The Open DB buttonis used when we want to read data from a database. The Generate button enables us to generate artificial data from a variety of DataGenerators. Using the Open File button we can read files in a variety of formats: ARFF, CSV, C4.5, or serialized instances format, these format have the extensions .arff, .csv, .data and .names, .bsi.

WEKA has converters for some file formats such as Spreadsheet files with extension .csv, C4.5’s native file format with extensions .names and .data, etc., This list of formats can be extended by adding custom file converters to the WEKA core converters package. The appropriate converter is used based on the file extension. If WEKA cannot load the data, it tries to interpret it as ARFF. If that fails, it pops up the generic object editor box, which is used throughout Weka for selecting and configuring an object. In this case the CSVLoader for .csv files is selected by default and the “more” button gives us more information about it. Figure 4.8 shows the Quran question& answer corpus file into Weka



Filter: None

Current relation
Relation: Quran question and anwer corpus
Attributes: 1
Instances: 30
Sum of weights: 30

Selected attribute
Name: question;answer
Type: Nominal
Missing: 0 (0%)
Distinct: 30
Unique: 30 (100%)

Attributes

No.	Name
1	question;answer

No.	Label	Count	Weight
1	On which night ...	1	1.0
2	Who revealed th...	1	1.0
3	Through whom ...	1	1.0
4	To whom was th...	1	1.0
5	Who took the re...	1	1.0
6	How many Surah...	1	1.0
7	What are the co...	1	1.0

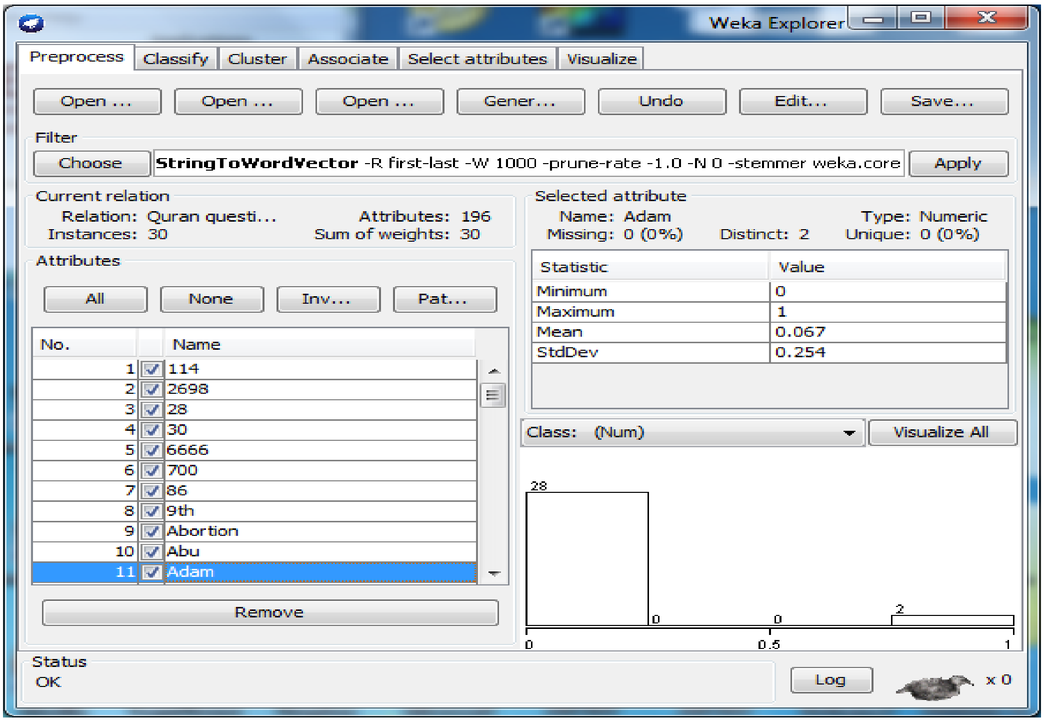
Class: question;answer (Nom)

Status
OK

Figure 4.8 the Quran question& answer corpus file loaded into Weka

4.3.1.3 Clustering the Quran Question and Answer Corpus

After loading the CSV file into WEKA’s Explorer, this dataset were processed into vectors of word frequencies using the StringToWordVector filter, which converts a string attributes to a “bag of words”, a vector that represents word occurrence frequencies. The StringToWordVector filter produces numeric attributes that represent the frequency of words in the value of each string attribute. The set of words (the new attribute set) is determined from the full set of values of all the strings in the full dataset. The list of all attributes, statistics and other parameters can be utilized as shown in Figure 4.9. There are 30 instances and 196 attributes in the “Quran question and answer” sample relation file.



Filter
StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core

Current relation
Relation: Quran questi... Attributes: 196
Instances: 30 Sum of weights: 30

Selected attribute
Name: Adam Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.067
StdDev	0.254

Attributes

No.		Name
1	☑	114
2	☑	2698
3	☑	28
4	☑	30
5	☑	6666
6	☑	700
7	☑	86
8	☑	9th
9	☑	Abortion
10	☑	Abu
11	☑	Adam

Class: (Num)

Status
OK

Figure 4.9the processed CSV file into WEKA

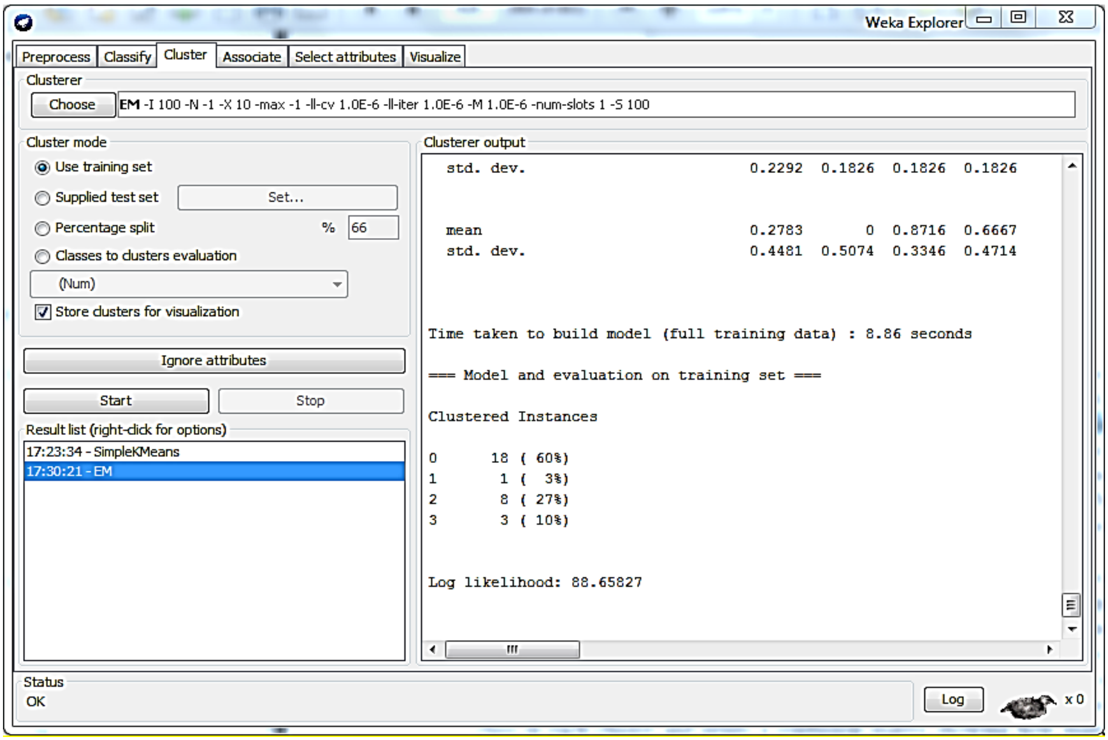
Clustering is used for data in which no class value is specified. In clustering, relevant attributes in the data are selected to decide the cluster. In some algorithms the number of clusters can be specified by setting a parameter in the object editor. For probabilistic clustering methods, WEKA measures the log-likelihood of the clusters on the training data: The larger this quantity, the better the model fits the data. Increasing the number of clusters normally increases the likelihood.

To cluster the Quran question and answer corpus, the ‘cluster’ pane was clicked to choose the requested clustering algorithms. We start with ‘SimpleKMeans’, the default number of cluster for SimpleKMeans algorithm is 2; but users can specify any number of clusters they want. K-Means clustering requires the user to input the desired number of clusters, but that might not be interesting as the user normally would not know in advance the number of clusters, and would be expecting the algorithm to cluster them into optimal groups. For this reason we used ‘expectation maximization’ or EM algorithms.

Our processed data were further analyzed using EM cluster. In EM clustering, the algorithm repeatedly refines an initial model to fit the data and find the probability that a data point exists in a cluster. The algorithm terminates the processing when the probabilistic model fits the data. The log-likelihood function is used to determine if the data were fit in the model. In case empty clusters are created during the process, or if one or more of the clusters have low populations that falls below a given threshold, then these clusters are reseeded at new points and the EM algorithm is rerun. The results of the EM clustering method are probabilistic. This means that every data point belongs to all clusters, but each assignment of a data point to a cluster has a different probability. Because the method allows for clusters to overlap, the sum of items in all the clusters may exceed the total items in the training set. The EM clustering algorithm offers multiple advantages in comparison to k-means clustering:

- Requires one database scan, at most.
- Will work despite limited memory (RAM).

The result ofEM algorithm inFigure 4.10 shows the attributes which are clustered, the number of clusters, and the number of instances each cluster contains. Figure 4.11shows WEKA cluster visualizer in which the attributes are clustered into 4 groups. The Visualize panel helps to visualize a dataset itself. It displays a matrix with a two-dimensional scatter plot.



Clusterer: EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -num-slots 1 -S 100

Cluster mode

- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation (Num)
- ☑ Store clusters for visualization

Result list (right-click for options)

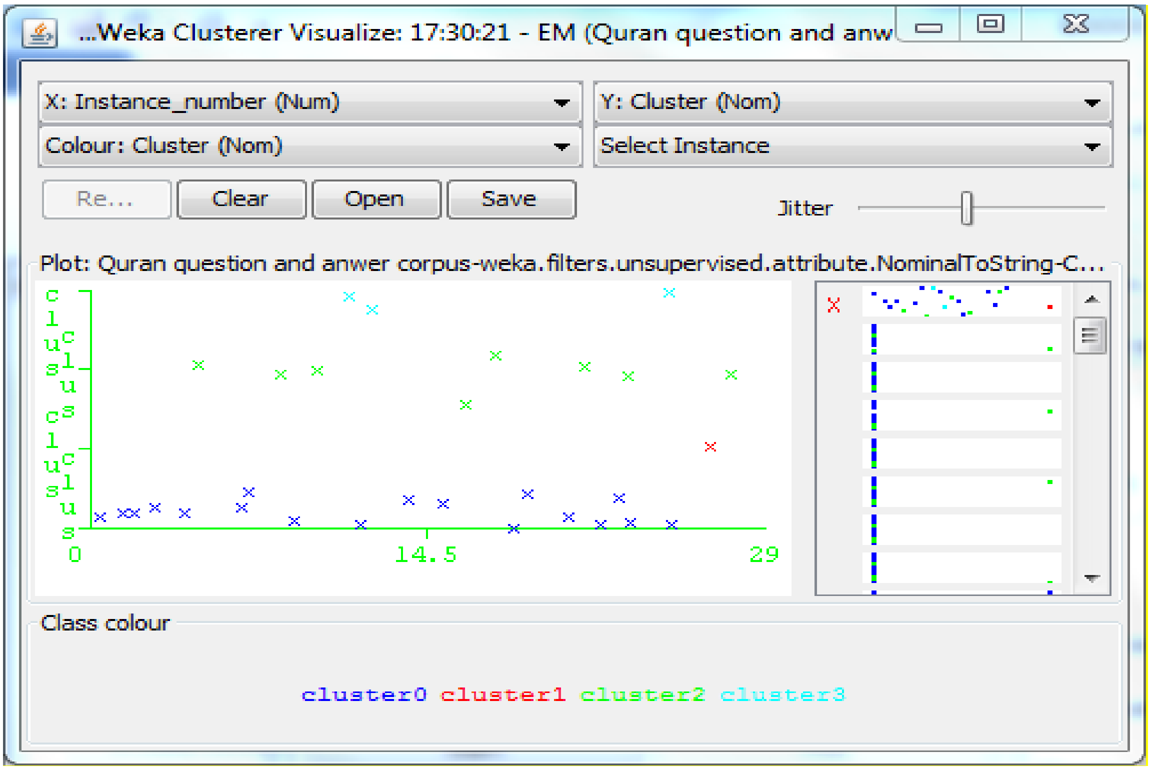
- 17:23:34 - SimpleKMeans
- 17:30:21 - EM

Clusterer output

```
  std. dev.                          0.2292  0.1826  0.1826  0.1826


  mean                               0.2783       0  0.8716  0.6667
  std. dev.                          0.4481  0.5074  0.3346  0.4714



Time taken to build model (full training data) : 8.86 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       18 ( 60%)
1        1 (  3%)
2        8 ( 27%)
3        3 ( 10%)


Log likelihood: 88.65827
```

Status
OK

Figure 4.10 the clusters and their instances



...Weka Clusterer Visualize: 17:30:21 - EM (Quran question and anw
Re...
Clear
Open
Save
Jitter
cluster
cluster
0
14.5
29

X: Instance_number (Num)

Y: Cluster (Nom)

Colour: Cluster (Nom)

Select Instance

Plot: Quran question and anwer corpus-weka.filters.unsupervised.attribute.NominalToString-C...

Class colour

cluster0 cluster1 cluster2 cluster3

Figure 4.11 visualizing the Quran questions and answers dataset

4.3.1.4 Results: WEKA

Figure 4.8 and Figure 4.9 show that 18 instances were clustered in cluster 0, 1 instance in cluster 1, 8 instances in cluster 2, and 3 instances in cluster 3. From Table4.4, it is evident that cluster 2 has questions of "how many" with number answers, and in cluster 3 the questions contain some of the same words, for example the words “name”, “prophet”, “mentioned”, and “Quran “ were found in the questions. Cluster 1 has questions containing words that did not appear in any other question like the words “Islamic”, “view”, and “Abortion”. Cluster 0 contains the rest of the questions.

We concluded that WEKA is good machine learning tool for classification, clustering etc.,but does not have functionality to build a question-answer interface and hence we continued to examine and explore NOOJ tool.

Table 4.4 example questions and answers in each cluster

Cluster no.	**Question**	**Answer**
Cluster 0	Who revealed the Quran?	Allah revealed the Quran
	On which night was the Quran first revealed?	LailatulQadr
	Through whom was the Quran revealed?	Through Angel Jibraeel
	Who took the responsibility of keeping the Quran safe?	Allah himself
	In which Surah (chapter) the	Surah al-Nessa

	law of inheritance is mentioned?	
	What are the conditions for holding or touching the Quran?	One has to be clean and to be with ablution
	Why do Muslims believe that the Prophet Muhammad is the final prophet?	Muslims believe that the Prophet Muhammad is the final prophet on the grounds that the Quran and hadith state so
	To whom was the Quran revealed?	To the last Prophet Mohammed
	Which is the longest Surah (Chapter) in the Quran?	Surah al-Baqarah
Cluster 1	What is the Islamic view on Abortion?	Islam considers abortion as murder and does not permit it
Cluster 2	How many verses are there in the Quran?	6666
	How many parts are there in the Quran?	30
	How many Makki Surah (chapter) are there in the Quran?	86

Cluster3	What is the name of the Prophet that mentioned and discussed most in the Quran?	Moosa (Alahis-Salaam)
	Who is the relative of the Prophet Mohammed (SallallahuAlaihiWasallam) whose name is mentioned in the Quran?	Abu Lahab

4.3.1.5 WEKA Errors while Loading Data

To load data into WEKA, we have to put it into a format that WEKA understands. WEKA needs the data to be present in ARFF or XRFF format in order to perform any tasks. When the format is incorrect while loading a certain file an error will occur; there are some reasons that caused that error for example wrong encoding file format or incompatible characters in the CSV Like a percentage sign (%), an apostrophe (‘), incorrect endings, and, any extra commas, etc.

4.3.1.6 Conclusions: WEKA

A representative sample of the question and answer corpus were selected to be used in WEKA model. Then the data were cleaned to improve data quality to the level required by the WEKA tool, and then converted to a comma separated value (CSV) file format to provide a suitable corpus dataset that can be loaded into WEKA. Then StringToWordVector filter was used to process each string into a bag or vector of word frequencies for further analysis with different data mining techniques. After that we applied a clustering algorithm to the processed attributes, and show the WEKA cluster visualizer.

We tried to move forward to develop a system that answers questions about the holy Quran using WEKA. We found that, WEKA is a good tool and very useful for classification and clustering etc., but does not have the quality of being suited well to serve our purpose in developing a question answering system for the holy Quran.

4.3.2 Using Nooj as an Analysis Tool for Quran Q&A Corpus

NOOJ is a free linguistic software that created by Max Silberztein in 2002. It processes text and corpus to build concordances; it is also used as information extractor for search engines, text mining and intelligence applications. NOOJ's architecture is based on the .NET technology; this allows NOOJ to work on documents on any computer, as well it allows other .NET applications to access NOOJ's public methods. NOOJ working on more than 100 file formats, including all different form of ASCII, Unicode, and HTML (Silberztein, Max, 2008). This section presents a NooJ module which aims to analyze, annotate and present the automatic processing of Quran's question and answers corpus.

4.3.2.1 Creating NOOJ file:

A CSV file containing a representative sample of 59 questions along with their answerswere selected from the collected questions and answers dataset to be imported to NOOJ. NooJ is usually used to work with text files that were created with another application. We need to import these files to Nooj in order to create a ".not" file and to parse them with NooJ. When we create a new NOOJ file we should have to perform a Linguistic Analysis before applying any queries and/or grammars to the text. Any

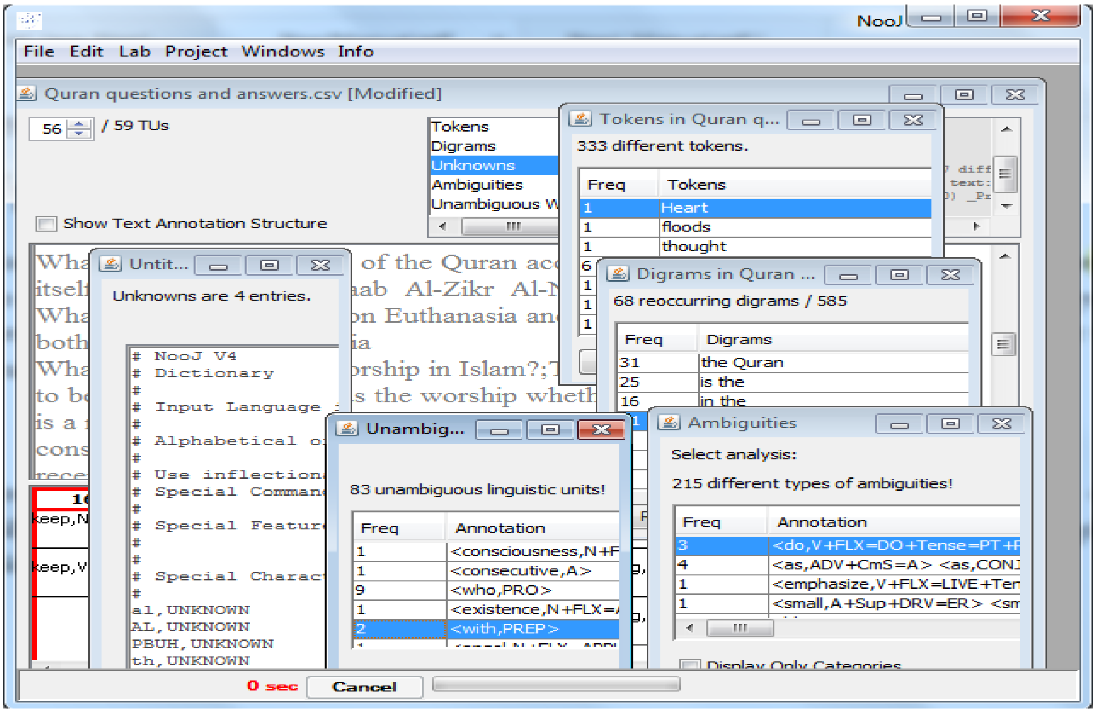
“.not” file can be modified, by using TEXT > Modify, as soon as we modify a text file, NOOJ erases all its Text Annotation Structure (TAS). Therefore, we need to perform a Linguistic Analysis again before applying any queries or grammars to it.

NOOJ processes the Text’s Units (TUs) one at a time. It is important to tell NOOJ what the text’s units are, because they control what exactly NOOJ can find in the text. Text Unit Delimiter that NOOJ uses are: (1) No delimiter: The whole text is seen as one large text unit, this option is useful when texts are small, and the information to be extracted is located everywhere in the text. (2)Text Units are lines/paragraphs: This is the default option. (3) PERL regular expression: This option allows users to define their own text unit delimiter. PERL regular expressions allow users to describe more sophisticated patterns. (4) XML Text Nodes: allows users to process structured XML documents. To import our CSV file in NOOJ, three parameters need to be set: we set the text’s language to “English”, file format to “ASCII or Byte-marked Unicode (UTF-8, UTF-16B, UTF-16L)”, and Unit Delimiters to ”Text Units are lines/Paragraphs”.

4.3.2.2 Applying Linguistic Analysis

After we imported the text file, we set the default working language, default fonts to display texts, as well as the lexical and syntactic resources that we want to apply to the Quran questions and answers text by using Info > Preferences control panel. Then the Tex > linguistic Analysis has been applied. The file has been saved in NOOJ’s format, because NooJ uses its own file format (.not) to process texts. Basically, “.not” NooJ text files store the text as well as structural information, various indices and linguistic annotations in the Text Annotation Structure. “.not” files, contains the text associated with linguistic information: usually, the text has been delimited into text units.

Each text unit is associated with a Text Annotation Structure, some dictionaries, morphological and syntactic grammars have already been applied to the text. As we can see in Figure4.12 as a result of linguistic analysis, NOOJ has produced some information, displayed above and to the right of the text window. Double click in the Results window, will display the lists of the text's Characters, Tokens, Diagrams, Unknowns, etc. Tokens are the basic linguistic objects processed by NOOJ; they are classified into three types: Word Forms, Digits; and Delimiters. Diagrams are pairs of word forms. The Annotations display the information that is being associated to the text, and the Unknowns display the word forms that have not been associated with any annotations (Max, Silberztein,. 2003) In the result we can see that there are 4 unknown entries.



Unknowns are 4 entries.

```
# NooJ V4
# Dictionary
#
# Input Language :
#
# Alphabetical o
#
# Use inflection
# Special Comman
#
# Special Featur
#
#
# Special Charac
#
al, UNKNOWN
AL, UNKNOWN
PBUH, UNKNOWN
th, UNKNOWN
```

333 different tokens.

Freq	Tokens
1	Heart
1	floods
1	thought

68 reoccurring digrams / 585

Freq	Digrams
31	the Quran
25	is the
16	in the

83 unambiguous linguistic units!

Freq	Annotation
1	<consciousness,N+F
1	<consecutive,A>
9	<who,PRO>
1	<existence,N+FLX=/
2	<with,PREP>

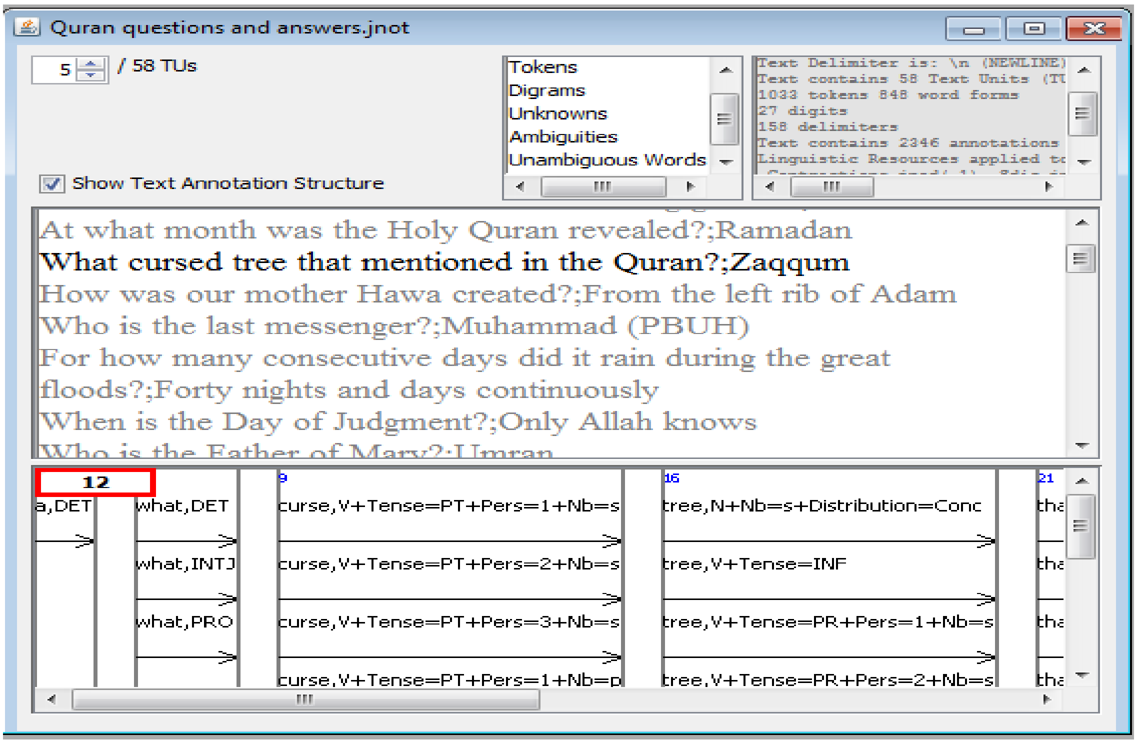
Select analysis:

215 different types of ambiguities!

Freq	Annotation
3	<do,V+FLX=DO+Tense=PT+F
4	<as,ADV+CmS=A> <as,CONJ
1	<emphasize,V+FLX=LIVE+Ter
1	<small,A+Sup+DRV=ER> <sm

Figure 4.12 the information produced as a result of linguistic analysis

Figure.4.13 shows NooJ's text annotation structure that gives the linguistic analysis of each word form in our sample question "What cursed tree that mentioned in the Quran? Zaqqum ". Annotation is a pair (position, information) that represents that a specific sequence of the text has certain properties. When NOOJ processes a text, it produces a set of annotations, stored in the Text Annotation Structure (TAS); annotations and original text file are always remains synchronized with each other's. Annotations contain information that can represent: lexical, morphological, syntactic, semantic, etc. NOOJ morphological and syntactic parsers possess tools to automatically add and remove annotations to and from a TAS.



Quran questions and answers.jnot

5 / 58 TUs

☑ Show Text Annotation Structure

Tokens
Digrams
Unknowns
Ambiguities
Unambiguous Words

```
Text Delimiter is: \n (NEWLINE)
Text contains 58 Text Units (TU
1033 tokens 848 word forms
27 digits
158 delimiters
Text contains 2346 annotations
Linguistic Resources applied to
```

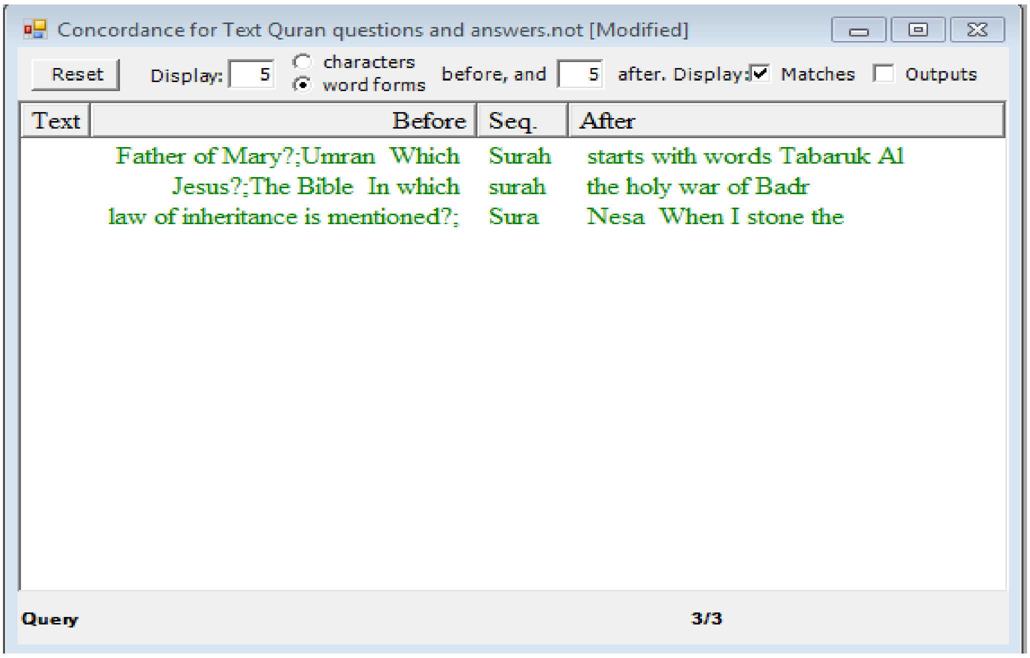
At what month was the Holy Quran revealed?;Ramadan
What cursed tree that mentioned in the Quran?;Zaqqum
How was our mother Hawa created?;From the left rib of Adam
Who is the last messenger?;Muhammad (PBUH)
For how many consecutive days did it rain during the great floods?;Forty nights and days continuously
When is the Day of Judgment?;Only Allah knows
Who is the Father of Mary?;Umran

```
12                  9                                16                                   21
a,DET   what,DET    curse,V+Tense=PT+Pers=1+Nb=s     tree,N+Nb=s+Distribution=Conc        tha
        what,INTJ   curse,V+Tense=PT+Pers=2+Nb=s     tree,V+Tense=INF                     tha
        what,PRO    curse,V+Tense=PT+Pers=3+Nb=s     tree,V+Tense=PR+Pers=1+Nb=s          tha
                    curse,V+Tense=PT+Pers=1+Nb=p     tree,V+Tense=PR+Pers=2+Nb=s          tha
```

Figure 4.13 the text annotation structure for a sample question

4.3.2.3 Using Regular Expressions to Retrieve Simple Patterns

The regular expressions are used in order to describe and retrieve simple morph-syntactic patterns in texts. To locate pattern in Quran question and answer data set, the TEXT >Locate Panel was used, then the option “a NooJ regular expression” was selected, after that the regular expression “surah | sura” was entered. To display the result we clicked one of the colored button in the window, NooJ let us know that it found 3 matches for our query, and then displayed a concordance in the selected color (green) as shown in Figure 4.14. The concordance of a sequence is an index that represents all of its utterances in context. NooJ concordances are displayed in four columns: each occurrence being presented in the middle column, between its left and its right context. If a corpus composed of a set of text files being indexed, rather than a single text file, then the first column displays the text file name in which each match occurs. It is also possible to apply a PERL regular expression, or a Nooj grammar or simply writing a string of character in the locate Panel’s text box.



Concordance for Text Quran questions and answers.not [Modified]

Reset Display: 5 ○ characters ● word forms before, and 5 after. Display: ☑ Matches □ Outputs

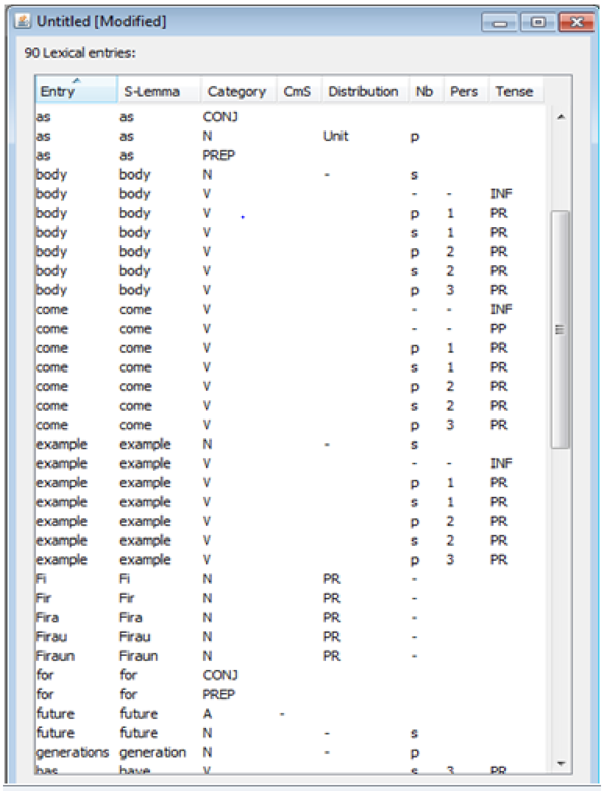
Text	Before	Seq.	After
	Father of Mary?;Umran Which	Surah	starts with words Tabaruk Al
	Jesus?;The Bible In which	surah	the holy war of Badr
	law of inheritance is mentioned?;	Sura	Nesa When I stone the

Query 3/3

Figure 4.14 the concordance for the query “Surah | sura”

4.3.2.4 Displaying Vocabulary for the Quran Q&A Corpus

In order to analyze texts, NOOJ needs dictionaries which, contain all words of a text and the description for these words, as well a mechanism to link these lexical entries to the corresponding inflected and/or derived forms that occur in texts. Dictionaries usually associate words or expressions with a set of information. NOOJ dictionary can include simple words, multi-word units, and can link lexical entries to a canonical form. NOOJ's dictionaries store syntactic and semantic information. Generally, the dictionary of a given language contains all of the lemmas of the language, and associates them with a morph-syntactical, possible syntactic and semantic codes, and inflectional and derivational paradigms. Figure15-a, and Figure.4.15-b shows a vocabulary which describes all of the words that comprises the Quran questions and answers in our CSV file, as a list and as table respectively.

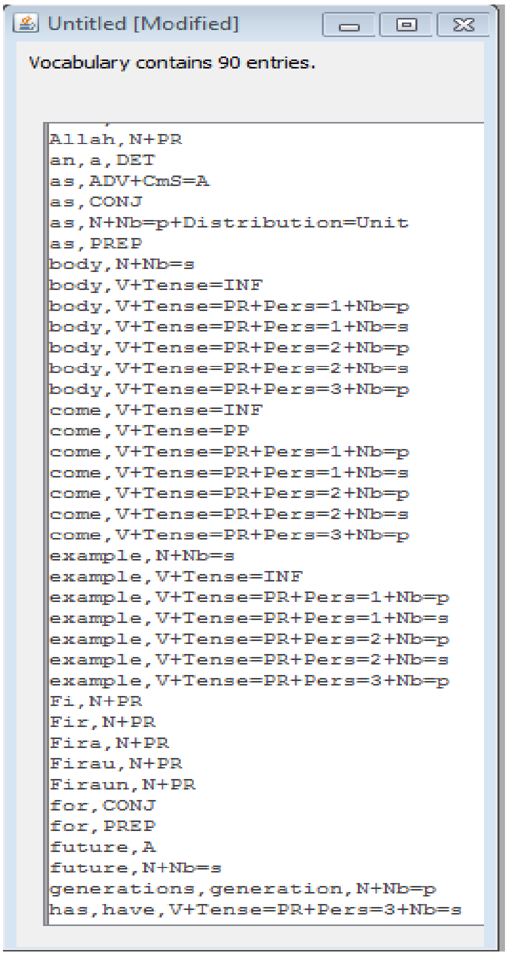


Untitled [Modified]

90 Lexical entries:

Entry	S-Lemma	Category	CmS	Distribution	Nb	Pers	Tense
as	as	CONJ					
as	as	N		Unit	p		
as	as	PREP					
body	body	N		-	s		
body	body	V			-	-	INF
body	body	V			p	1	PR
body	body	V			s	1	PR
body	body	V			p	2	PR
body	body	V			s	2	PR
body	body	V			p	3	PR
come	come	V			-	-	INF
come	come	V			-	-	PP
come	come	V			p	1	PR
come	come	V			s	1	PR
come	come	V			p	2	PR
come	come	V			s	2	PR
come	come	V			p	3	PR
example	example	N		-	s		
example	example	V			-	-	INF
example	example	V			p	1	PR
example	example	V			s	1	PR
example	example	V			p	2	PR
example	example	V			s	2	PR
example	example	V			p	3	PR
Fi	Fi	N		PR	-		
Fir	Fir	N		PR	-		
Fira	Fira	N		PR	-		
Firau	Firau	N		PR	-		
Firaun	Firaun	N		PR	-		
for	for	CONJ					
for	for	PREP					
future	future	A	-				
future	future	N		-	s		
generations	generation	N		-	p		
has	have	V			s	3	PR

Figure 4.15-a dictionary describes all words as a table



Vocabulary contains 90 entries.

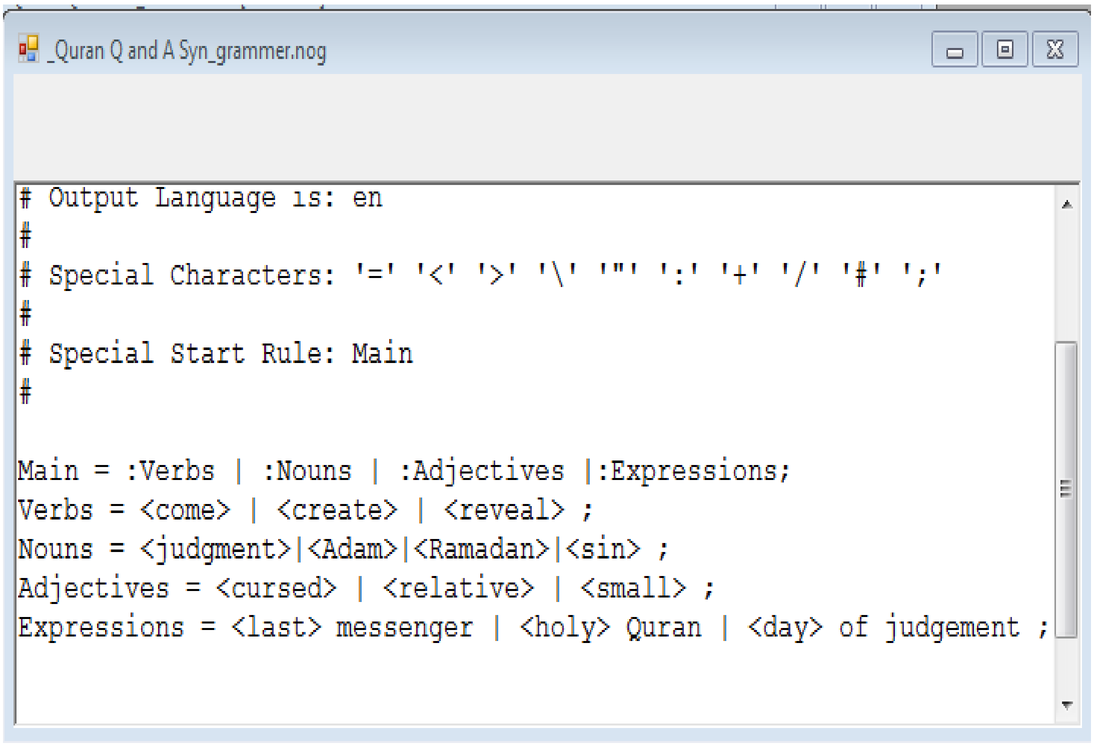
```
Allah,N+PR
an,a,DET
as,ADV+CmS=A
as,CONJ
as,N+Nb=p+Distribution=Unit
as,PREP
body,N+Nb=s
body,V+Tense=INF
body,V+Tense=PR+Pers=1+Nb=p
body,V+Tense=PR+Pers=1+Nb=s
body,V+Tense=PR+Pers=2+Nb=p
body,V+Tense=PR+Pers=2+Nb=s
body,V+Tense=PR+Pers=3+Nb=p
come,V+Tense=INF
come,V+Tense=PP
come,V+Tense=PR+Pers=1+Nb=p
come,V+Tense=PR+Pers=1+Nb=s
come,V+Tense=PR+Pers=2+Nb=p
come,V+Tense=PR+Pers=2+Nb=s
come,V+Tense=PR+Pers=3+Nb=p
example,N+Nb=s
example,V+Tense=INF
example,V+Tense=PR+Pers=1+Nb=p
example,V+Tense=PR+Pers=1+Nb=s
example,V+Tense=PR+Pers=2+Nb=p
example,V+Tense=PR+Pers=2+Nb=s
example,V+Tense=PR+Pers=3+Nb=p
Fi,N+PR
Fir,N+PR
Fira,N+PR
Firau,N+PR
Firaun,N+PR
for,CONJ
for,PREP
future,A
future,N+Nb=s
generations,generation,N+Nb=p
has,have,V+Tense=PR+Pers=3+Nb=s
```

Figure 4.15-b dictionary describes all words as a list

4.3.2.5 Creating a Grammar

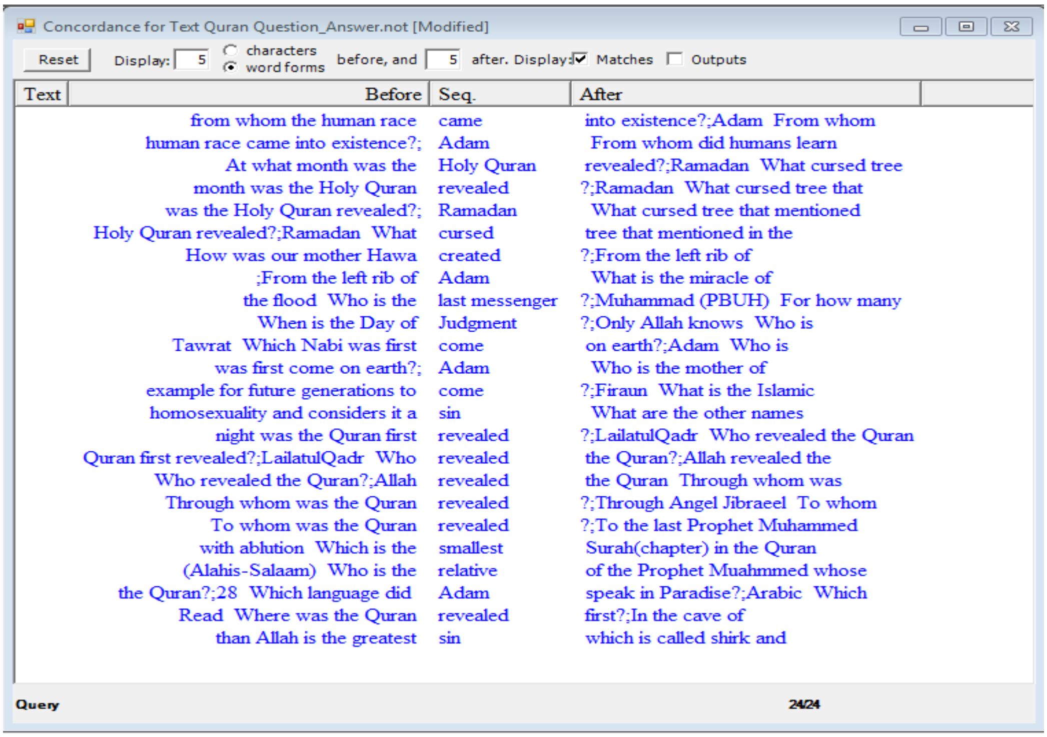
Grammar is used to represent a large spectrum of linguistic phenomena, it is used to extract sequences of interest in texts, and to describe various linguistic phenomena. Grammars can contain a large number of embedded graphs. In NOOJ; there are different types of grammars: (a) Inflectional and derivational grammars (.nof files) are used to represent the inflection or the derivation properties of lexical entries. These descriptions can be entered either graphically or in the form of rules. (b) Lexical, orthographical, morphological or terminological grammars (.nom files) are used to represent sets of word forms, and associate them with lexical information, (c) Syntactic, semantic and translation grammars (.nog files) are used to recognize and annotate expressions in texts.

All types of grammars are represented by organized sets of graphs. A grammar is a collection of one or more queries; each query has a name, and must be ended by a semi-colon. We can store any numbers of queries in a single grammar file. Each NooJ grammar contains only one special rule named Main, each query may contain references to other queries. We tried to create a small grammar to locate some specific pattern in our corpus; Figure4.16 shows that the rule “Main” refers to the four rules Verbs, Nouns, Adjectives and Expressions Then we applied the created grammar to our text to display the concordance as shown in Figure4.17. Nooj’s grammars are more powerful than regular expressions.



```
# Output Language is: en
#
# Special Characters: '=' '<' '>' '\' '"' ':' '+' '/' '#' ';'
#
# Special Start Rule: Main
#

Main = :Verbs | :Nouns | :Adjectives |:Expressions;
Verbs = <come> | <create> | <reveal> ;
Nouns = <judgment>|<Adam>|<Ramadan>|<sin> ;
Adjectives = <cursed> | <relative> | <small> ;
Expressions = <last> messenger | <holy> Quran | <day> of judgement ;
```

Figure 4.16 the “Main” rule composes of four rules Verbs, Nouns, Adjectives and Expressions.


Concordance for Text Quran Question_Answer.not [Modified]
Reset
Display: 5
○ characters
● word forms
before, and 5
after. Display: ☑ Matches
□ Outputs
Query
24/24

Text	Before	Seq.	After
	from whom the human race	came	into existence?;Adam From whom
	human race came into existence?;	Adam	From whom did humans learn
	At what month was the	Holy Quran	revealed?;Ramadan What cursed tree
	month was the Holy Quran	revealed	?;Ramadan What cursed tree that
	was the Holy Quran revealed?;	Ramadan	What cursed tree that mentioned
	Holy Quran revealed?;Ramadan What	cursed	tree that mentioned in the
	How was our mother Hawa	created	?;From the left rib of
	;From the left rib of	Adam	What is the miracle of
	the flood Who is the	last messenger	?;Muhammad (PBUH) For how many
	When is the Day of	Judgment	?;Only Allah knows Who is
	Tawrat Which Nabi was first	come	on earth?;Adam Who is
	was first come on earth?;	Adam	Who is the mother of
	example for future generations to	come	?;Firaun What is the Islamic
	homosexuality and considers it a	sin	What are the other names
	night was the Quran first	revealed	?;LailatulQadr Who revealed the Quran
	Quran first revealed?;LailatulQadr Who	revealed	the Quran?;Allah revealed the
	Who revealed the Quran?;Allah	revealed	the Quran Through whom was
	Through whom was the Quran	revealed	?;Through Angel Jibraeel To whom
	To whom was the Quran	revealed	?;To the last Prophet Muhammed
	with ablution Which is the	smallest	Surah(chapter) in the Quran
	(Alahis-Salaam) Who is the	relative	of the Prophet Muahmmed whose
	the Quran?;28 Which language did	Adam	speak in Paradise?;Arabic Which
	Read Where was the Quran	revealed	first?;In the cave of
	than Allah is the greatest	sin	which is called shirk and

Figure 4.17 the concordance for the created grammar

4.3.2.6 Importing and Exporting Documents to and from Nooj

NOOJ can import an XML document, and automatically converts XML tags to NOOJ annotations. The ability to import XML tags as NOOJ's lexical or syntactic/semantic annotations allows NOOJ to parse texts that have been processed with other applications. NOOJ can also export a Text Annotation Structure as an XML document by using TEXT > Export annotated text as an XML document. NOOJ's annotations will be represented as XML tags and inserted in the resulting text and can be used by other applications

4.3.2.7 Conclusions: Nooj

NooJ is a linguistic development environment that allows users to create formalized dictionaries and grammars and use these resources to build some NLP applications. NooJ allows users to process large sets of texts.

In this section, we explored the use of Noojto process our Quran Q&A Corpus. We found that Nooj offered some analysis methods but Nooj does not offer enough tools for question answering system development. So, we had to experiment with another toolkit instead.

4.3.3 Python &Natural Language ToolKit (NLTK)

Python is a high-level programming language; developed by Guido van Rossum and first released in 1991. It is a widely used general purpose programming language. Python is an interpreted and object-oriented programming language. It possesses strong build-in libraries which make it simple and easy to learn and use. Python is regarded as platform independent language; the program created in python can be run on different platform and under different operating systems. The syntax and the semantics of python are very clear and concise; it has philosophy in its design which emphasizes code readability where it uses whitespace indentation to delimit code blocks rather than curly braces or keywords, and a syntax which allows programmers to express concepts in fewer lines.

Python's NLTK is a leading platform for building Python programs to work with symbolic and statistical natural language processing. It is a comprehensive Python library for natural language processing and text analytics (Perkins, J.2014), such as classification, tokenization, stemming, tagging, parsing, and semantic reasoning. NLTK includes graphical demonstrations and sample data. It provides easy-to-use interfaces to more than 50 corpora and lexical resources such as WordNet. It was developed by Steven Bird and Edward Loper in 2001 as part of a computational linguistics course in the Department of Computer and Information Science at the University of Pennsylvania. Since then it has been developed and expanded with the help of many contributors. It has now been used as the basis of many research projects. (Bird, S., et al, 2014)

4.4 Developingan Automated QASystem forthe Holy Quran

4.4.1 Introduction

Question Answering (QA) systems can be defined as automated systems capable of accepting a question posed in natural language, processing this question, and returning the required information as an answer to the asked question. This thesis proposes to compile QAEQ&AC (Quranic Arabic/EnglishQuestion and Answer Corpus) as knowledge base and applying existing advanced information retrieval techniques, and natural language processing methods to create a QA system for the holy Quran, which is called QAEQAS: Quran Arabic/English Question Answering System. As a complete question answering solution, we used the python and toolkit to process the user question and the corpus as well as to implement the search engine to retrieve candidate results and then extract the best answer.

A question answering system relies on a good search corpus: If documents do not contain the answer the system cannot do more. If the answer to a question is not existent in the data sources, a correct answer will not be obtained, even if the question processing, information retrieval and answer extraction models are doing well. It is thus clear that larger collection sizes of corpus normally deliver better QA performance; this can be done by integrating an enormous range of knowledge-bases. The greater the number of sources of information from which we extract the required results, the easier the task becomes, because the required result can be expressed in different ways. The data redundancy in huge collections means that some of the information may be phrased in many different ways in differing contexts, thus the onus on the QA system to perform complex NLP techniques to understand the text may be reduced because the right information appears in many forms. Our system is different from most question answering systems that it depends on a knowledge base composed of question-answer pair and data redundancy instead of using complicated linguistic analyses.

In the previous section we explained all the steps to create the corpus for the holy Quran. This corpus is composed of a set of frequently asked questions along with their correct answers, which have been collected from several trusted expert sources, and then pre-processed. Different data subsets were merged for the purpose of a Quran Q&A task. The corpus collection comprises new data and some existing data from various small test-sets.

This section consists of three main ideas. Users are allowed to type a question in natural language throw a graphical user interface (GUI); the system will match the input with the stored corpus questions to extract the semantically similar question and then find the attached answer for this question. This is the principle of a FAQ answering system, which has been widely used in different manner and purpose and becoming an important component of QA system.

4.4.2 System Architecture

The general architecture of our question answering system is presented in Figure 4.18. It consists of four main modules: 1- Pre-processing module, 2- Information Retrieval module, 3- Scoring and Ranking Module, 4- Answer extraction Module.

Pre-Processing Module: It is a natural language analyzer module that includes mainly four operations:

- Normalization
- Removing the punctuations
- Tokenization
- Removing the stop words

Information Retrieval Module: It is a word matching module, the main target of this module is to provideaset of candidate questions that includes the question similar to the user question and return the results to the ranker module. This module includes two operations:

- Filtering questions from answers.
- Searching for candidate questions

Scoring and Ranking Module: The scoring and Ranking Moduleidentifies the closest question of the candidate questions that matches the question asked by user, by assigning scores tothe candidate questions which are retrieved from IR module, and then ranking the candidate questions in descending order.

Answer Extraction Module:The Answer Extraction Module finds the best candidatequestion according to their scoresand then extract its answer as an answer for the user’s question

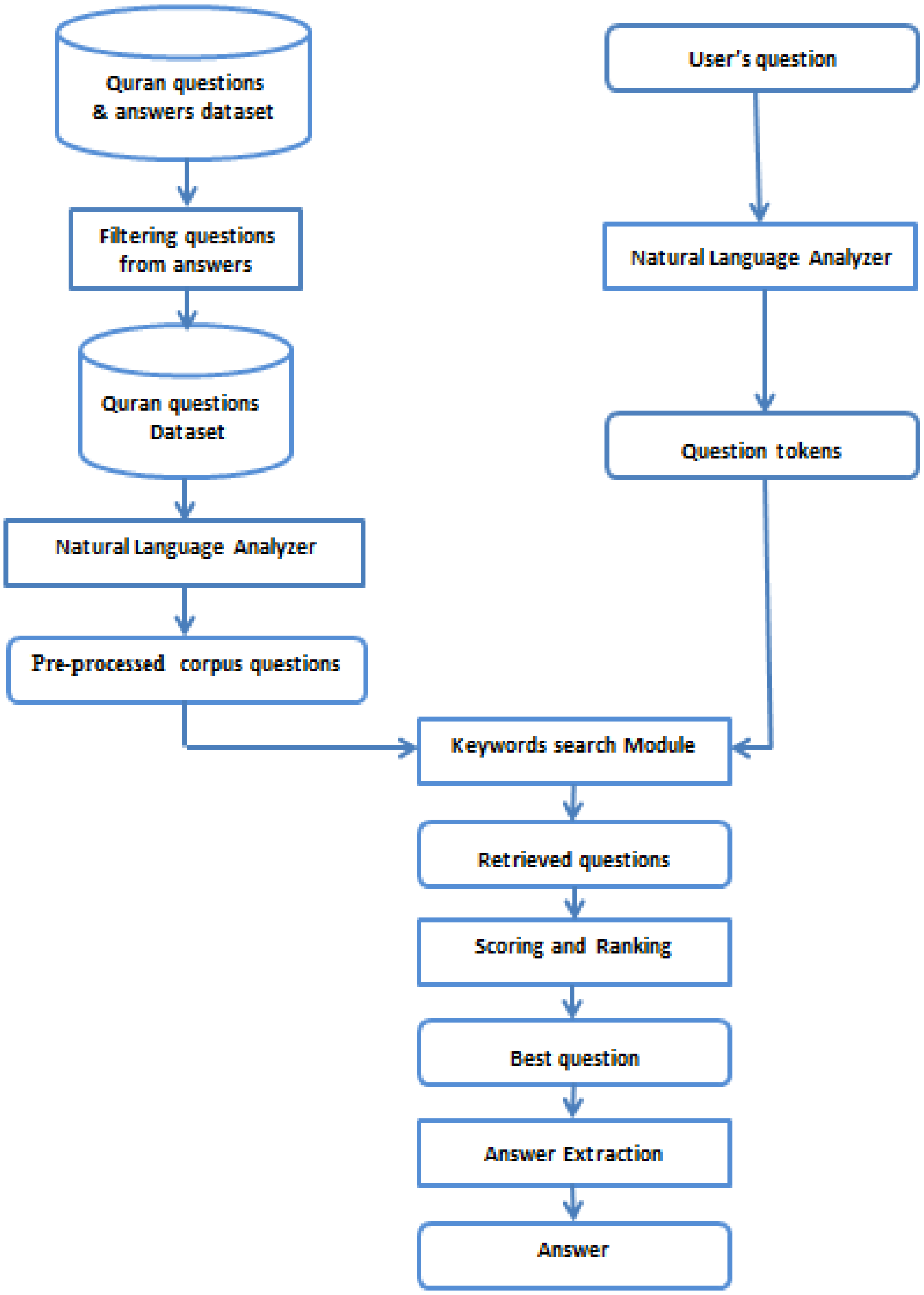


Figure 4.18 the general architecture ofQAEQAS

4.4.3 System Model

One experiment has been conducted for each version of the prototype. The first version used three variants of each question. While in the second version the question redundancy has been increased by using six variants of each question. This means that the redundancy for the second prototype is two times compared with the first prototype. This redundancy has been done by manually reformulating each question in different context to be semantically similar but syntactically different to the original question.

4.4.3.1 GeneralDescription

In our prototype versions to answer questions about the holy Quran, a single source of data has been used as knowledge base for each of Arabic and English version. Keyword-based techniques have been applied to return an answer. The general idea of QAEQAS is: In the input stage the system takes and accepts a Natural Language (NL) question from the user through a graphical user interface. Then matches this question with the questions found in the data set. In the outputstage the system returns the answer for the matched question as an answer for the user question. The answers may be short phrases or longer answers, like sentences, or even many paragraphs, to provide for question clarification.

As a complete question answering solution, the Python and natural language toolkit (nltk) was used to process the user question and the corpus, as well as to work as search engine to retrieve candidate results and finally extract the answer. While the context of the input question can differ from user to user, some of the questions were phrased in different ways in differing contexts. Data redundancy in large collections makes it easier for the system to find the right information. QAEQAS deals with a wide range of question types including: facts, definitions, how, why, etc.

4.4.3.2 Preparing the Knowledge Base File

Some questions along with their answers have been selected from the collected dataset, which were used as examples for the model. This dataset is composed of 500 questions along with their answers for the Arabic and

350 for the English. Then some of these questions were manually reformulated so that each question has two extra variant in different context and then added to the same file to be used in the first prototype version. The same work has been done for the second version, but this time each question was reformulated into up to sixvariants. Since the Excel application produces and uses CSV files, these questions and their answers for each language were entered in one sheet of an Excel workbook file, so that each question and its corresponding answer is a record (row). The table consists of two column with heading “question” and “answer”. After that the excel file has been converted into the appropriate comma-separated value (CSV) format which is convenient and suitable for our question answering system.

4.4.3.3 Building a Graphical User Interface

Python tkinter module was used for creating a graphical user interface (GUI) that facilitates user interaction with QAEQAS. The GUI allows a user to enter his/her question in natural language, and accepts the question. After the system processes the question and searches for the answer, the GUI accepts the returned answer if there is any, or a message telling the user that there is no answer for the question. Python has more than one GUI package, but tkinter is the standard and the most commonly used one. The tkinter module (Tk interface) is also the standard Python interface to the toolkit GUI from scriptics. tkinter comprises of a number of modules. The Tk interface is provided by a binary extension module named _tkinter. The public interface is provided through a number of Python modules. To use tkinter, we should import the tkinter module using: import tkinter, or, from Tkinter import * (Lundh, 1999).

4.4.3.4 Pre-processing Module

The most common pre-processing tasks in natural language processing are: Shifting everything to lower case, normalization, stripping punctuation, tokenization, and removing stop words. First the user question and the corpus questions were converted to lower case for English version, then the following steps has been done:

Normalization of Arabic Text

For a number of purposes Arabic text must be normalized, namely noise characters should be deleted, the orthography of problematic letters should be unified, etc.The Python's re.sub() function was used to unify the orthography of alifs, hamzas, and yas/alifmaqsuras, considering the probabilities that the user may use to writes Arabic characters, e.g., he may write "Alif" character without " alifs, hamzas"; or with alifs, hamzas., above or below the Alif. The short vowels and other symbols (harakat) that interfere with computational manipulations with Arabic text were also handled using re.sub() method. One of the most important re methods that use regular expressions is sub. This method substitutes all occurrences of the RE pattern in string with new ones and returns modified string.

Removing the Punctuations

Punctuation marks are symbols that are used in writing to separate sentences and their elements to aid the clarity and comprehension of written language. Some commonly used punctuation marks in Arabic and English grammar are: the period, question mark, exclamation point, comma, semicolon, colon, dash, hyphen, parentheses, brackets, braces, apostrophe, and quotation marks. What punctuations that should be removed aredepending on the nature of the data, and the task that should be done. Before tokenizing the user's question, the punctuations that will probably affect the processing were removed.

Tokenizing the User Question

Tokenization is the process of segmenting a text into linguistic units such as words, punctuation, numbers, alphanumeric, symbols, phrases, paragraphs, sentences or any other expressive elements, which known as tokens. This tokenized text is used as an input for further processing. Python Natural Language Tool Kit NLTK provides a number of tokenizers in the tokenize module, which are used to split strings into lists of substrings (Bird et al, 2009). They provide several ways to tokenize text: word_tokenizetokenizer is used to detect words and punctuation in a text, it needs the Punkt sentence tokenization models to be installed.

NLTK also has a simple tokenizer known as wordpunct_tokenize, used to divide text on whitespace and punctuation. wordpunct_tokenize is based on regular expressions. Tokenization can also work at the level of sentences. Using the sentence tokenizer by importing sent_tokenize, word_tokenize to divide the text to sentences and then to words (Perkins, 2014). One extra problem with Arabic questions and answers is that these tokenizers work on decoded version of a Unicode string only and not encoded version, therefore with Arabic text the string should be decoded first. To decode a string in Python we use the s.decode("utf8"). The word_tokenizer(s) was used to split the user question into tokens composed of words and punctuation.

Removing the Stop Words from the User Question

Stop words are the most common words in any language such as ("the", "a", or "and") in English language, which help build ideas, for example, linking sentences or words, but do not give any meaning. Stop words are words that we decide not to index at all, and therefore do not contribute in any way to retrieval and scoring. These words probably appear in many questions, so they are often separated out and removed before or after the processing of natural language data (text). There is no specific standard list of stop words that is used by all natural language processing tools. Any collection of words can be selected as the stop words for a specific purpose, stop words differ according to the case used.

It is clear that the words in a document (question) are not equally important;some of the most common words within a corpus do not play a considerable role in processing, such words are examples of domain-specific stop words. During processing, the presence of these words may give misleading results. There is no definite list of stop words as different systems use different stop words according to their requirements. For example some search engines remove some of the most common words — including lexical words, such as "want"— from a query in order to improve performance.

Our methodology used the English stop words list that included within NLTK, and a version of Arabic stop words list created by TahaZerrouki (Zerrouki et al 2014)as Arabic stop words list are not included in NLTK. As well assomedomain-specific words which, frequently appear in the Quran Q & A corpus were added to the stop word list for both Arabic and English. Then all these stop words were removed from the user question to minimize their effect and hence improve the performance of the system.

4.4.3.5 Information Retrieval Model

The mission of the IR module is to perform a first selection of documents (questions) that are considered relevant to the input question. The selection of an adequate retrieval model that fit the specific characteristic of the supplied data is considered as a core part of the task. To find similar questions and extract their corresponding answers. Priorto preprocess the corpus, the questions have been filtered from their answers using Python's split() method. Python provides a very straightforward and easy function to split or breakup a string and adds the data to a string array using a defined separator. If no separator is defined when we call upon the function, whitespace will be used by default. In simpler terms, the separator is a defined character that will be placed between each variable.

As soon as the QA system receives a question from the user, it first retrieves a set of candidate questions from the Quran questions. The regular expressions (re) module in Python provides full support regular expression matching operations. Regular expressions are a powerful and flexible mechanism of specifying patterns. In order to find matches among user question and corpus questionswe used regular expressions as they are very effective for string matching. Regular expression is special sequence of characters that helps to match or find other strings or sets of strings using a specialized syntax held in a pattern. A pattern is simply one or more characters that represent a set of possible match characters.

Regular expressions are patterns that permit us to "match" various string values in a variety of ways. In regular expression matching, we use a character (or set of characters) to represent the strings that we want to match in the text. The "re" package is used to carry out queries on an input text file. It has several functions such as re.match(), re.search() and re.findall(). Each of these functions takes a regular expression, and string to find matches. The match function only finds matches if they exist at the beginning of the string. The re.search function scans forward through string looking for a position where the regular expression pattern finds a match, and then returns a corresponding matched result. Both functions return "None" if no match can be found. To retrieve candidate questions the re.search() function has been used to search the corpusquestions. This is done by comparing all terms of the user question, one by one, with every corpus question to return the candidate matched questions. The regular expression "re.I" can be used to enable a case-insensitive mode (ignoring case), for example, "surah" will match the string "Surah".

First, the corpus has been searched without doing any preprocessing for the user question, in this case no answer has been found. Then after tokenizing the user question and without removing the stop words, the result was too many matches. But after removing the stop words there were fewer matched displayed than before. After that this group of stop words has been studied and examined against the corpus, in both English and Arabic, and we realized that there are some words that appear in many questions of the corpus, but were not included within the standard stop words lists. Some of these words are used to help in building questions such as the Arabic words: (رأى، وصف، ذكر، ورد),and English words (mention, describe etc.). Other words are considered as common words in Quran domain or might be used frequently in Islamic literature such as Arabic words, (سبحانه، تعالى، عزوجل، الكريم، (ص)), and English words (Peace, upon, Sallallahu, Alaihi, Wasallam, holy, etc.). Such words should be added to the stop words lists to improve the search performance. After adding this word to the stop word list and removing the stop words from the user question very fewer matched questions was displayed

4.4.3.6 Scoring and Ranking Model

As it has been seen before in the information retrieval model, QAEQAS displays a collection of candidate results, but we don’t know which the best ones are. The main task of the scoring and ranking model is to select the more relevant question from the candidate results in order to return the best answer. One possibility is to use a similarity measure to rank the candidates. The ideal ranking method for this task should be adaptable enough to fit the specific characteristics of the data in use. One good feature is the degree of match between the user query and each of the candidate results.

The score for thecandidates’ result that were returned after searching were used for this purpose. It is essential for python’s search engine to rank-order the candidate results matching a query. To do this, the search engine computes, for each candidate result, a score with respect to the query at hand. A score was then given to each of these candidates according to the number of question words it contains, the more the better. Assigning scores represents the degree of results relevance in respect to the user question.

It has been noticed that sometimes the same word may appear more than once in the candidate and gives it increased scores. This problem has been handled by counting the repeated word only once in the score. To do this the python set() method is used to allow no duplicate words.The score for each candidate question was calculatedusing the intersection among words between candidate question and the user question. After that the results have been ranked in descending order according to their score to obtain the best question. The python’s sorted ( ) built-in function was used for this purposein order to build a new sorted list. There are many ways to use this function to sort data, sorted ( ) functions accept a reverse parameter with a Boolean value. This is using to flag descending sorts. It has been used to get the candidate questions in reverse order according to their score.

4.4.3.7 Answer Extraction Model

The last task in question answering system is answer extraction. Answer extractionis one of the main tasks for QA systems in order to return the best answer. The input for the Answer Extraction Model is the ranked candidate questions, which produced by Scoring and Ranking Model. In our prototype the answer extraction process relies on extracting the best question -which matches the user question- from the ranked candidate questions, and then uses this question to find the complete line which is composed of the question along with its answer. After obtaining the line, Python's split() function have been used to divide the line into two partsnamely the question and the answer, then the answer for this similar question will be display as an answer for the user question.

4.4.4 QAEQAS with More Data Redundancy

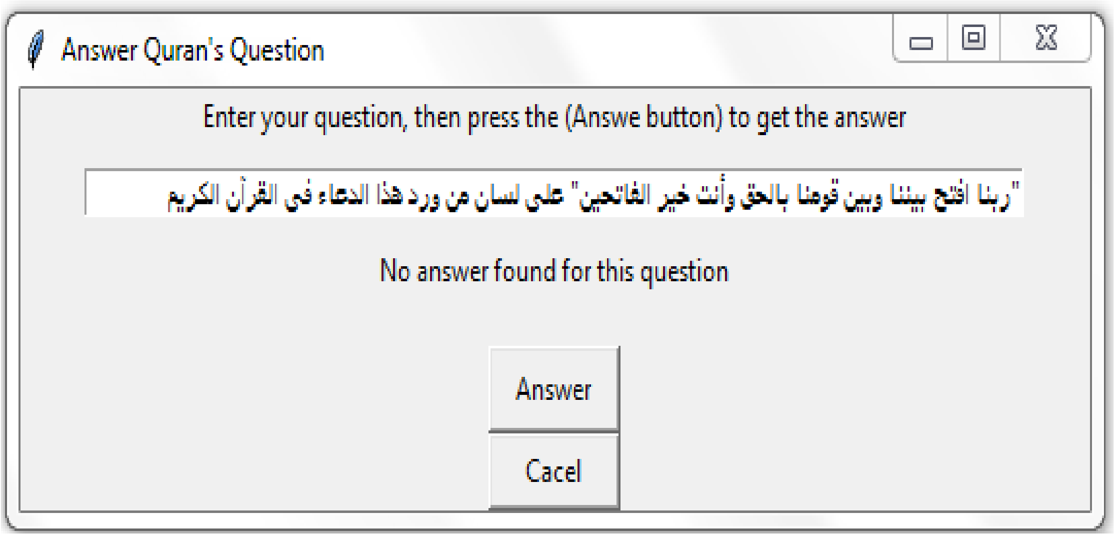
As we have mentioned before the greater the question redundancy in the source, the more likely it is that we can find a question in the corpus that semantically matches the user question. As well as the larger the data set from which we can draw answers, the greater the chance we can find an answer. Our approach to question answering is to take advantage of data redundancy, since the same question can be asked many times but in different context. In our methodology the data redundancy in huge collections means that some of the information may be phrased in many different ways in differing contexts.

Both versions of the prototype were benefited from the data redundancy that already found in the corpus due to data collection from different sources, hence the same question can be found in different context. In addition to that more data redundancy has been used in both versions of the prototype by manually reformulating some of the corpus questions into different contexts to be similar in semantic sense but different in syntactic sense, and then added to the file of the Quran question and answer corpus.

In the first version of the prototype each question was found into three differentvariants, the original question plus two extra variant. In the second version of the prototype the data redundancy has been increased two times in comparison to the first version, by reformulating each question into up to six different variants. Both versions of the prototype were tested using a set of questions asked by expert user group of Muslim university academics. Different results were obtained; the accuracy of the system was more increased by increasing the data redundancy.

4.4.5 Results: QAEQAS

Searching for answers in questions & answers corpus (knowledge base) without tokenizing the user's question gives no relevant result as shown in Figure.4.19 and Figure 4.20, because the system searches for the whole question. It gives an answer only when the user question is a full exact match to a question found in the corpus, and that may happened rarely.

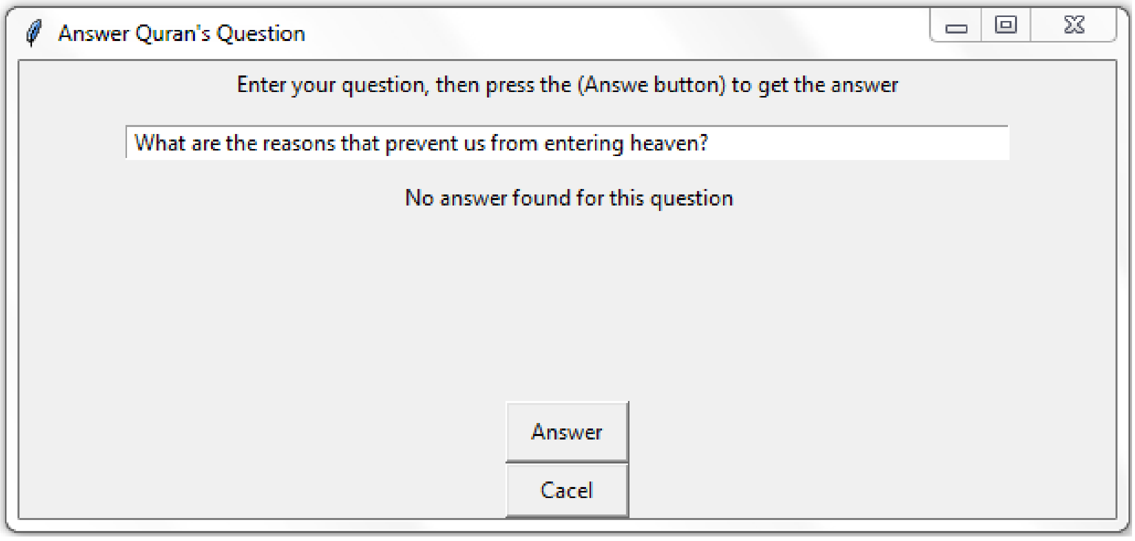


Enter your question, then press the (Answe button) to get the answer

"ربنا افتح بيننا وبين قومنا بالحق وأنت خير الفاتحين" على لسان من ورد هذا الدعاء في القرآن الكريم

No answer found for this question

Figure.4.19 the search result before tokenizing an Arabic user question



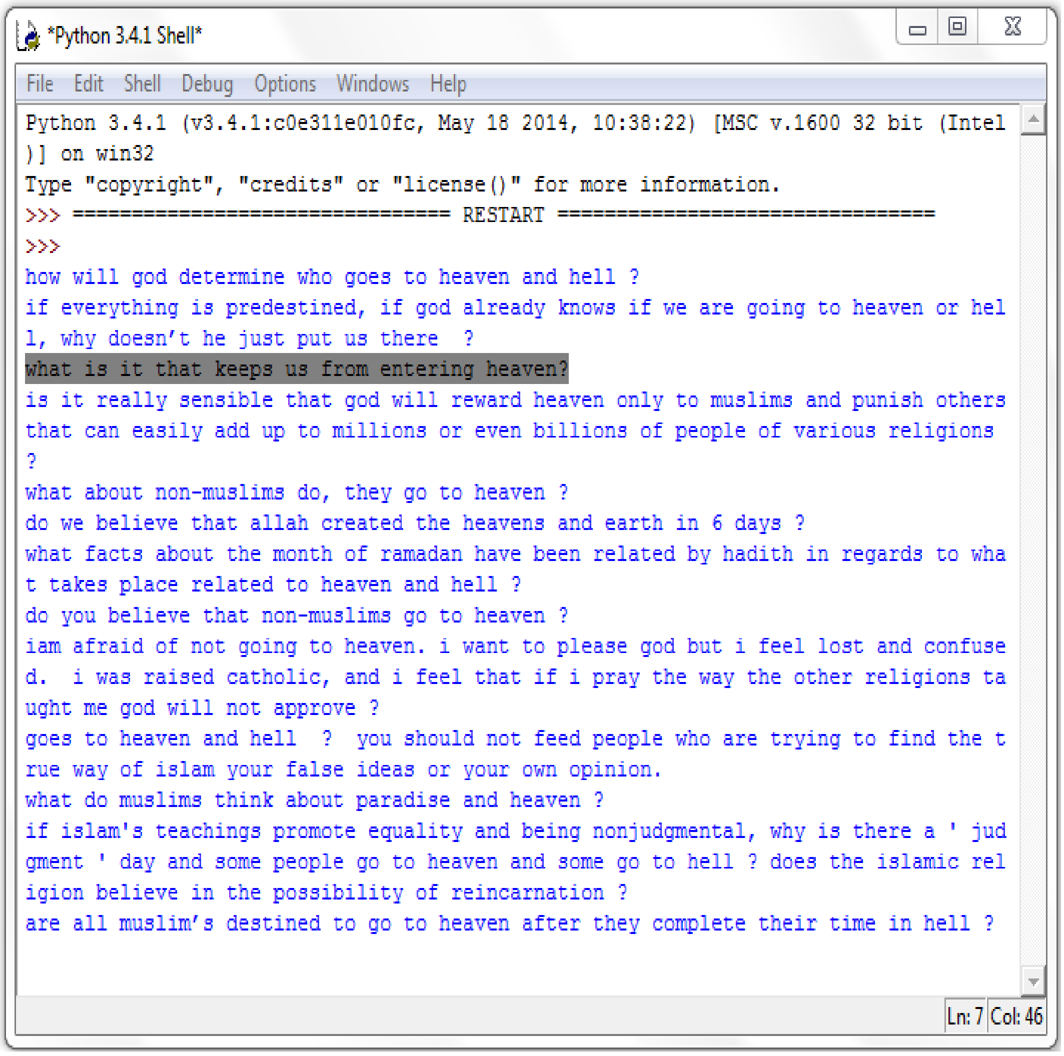
Enter your question, then press the (Answe button) to get the answer

What are the reasons that prevent us from entering heaven?

No answer found for this question

Figure.4.20 the search result before tokenizing an English user query

After tokenizing the user question, many answers have been displayed, because the system searches for matches in the corpus questions for every keyword found in the user question. After removing standard stop words the numbers of answers became less than before. When adding some words to the standard stop words and removing other stop words, the system displayed even fewer answers; Figure 4.21 shows the candidates result for the user' question "What are the reasons that prevent us from entering heaven?", after removing all stop words.

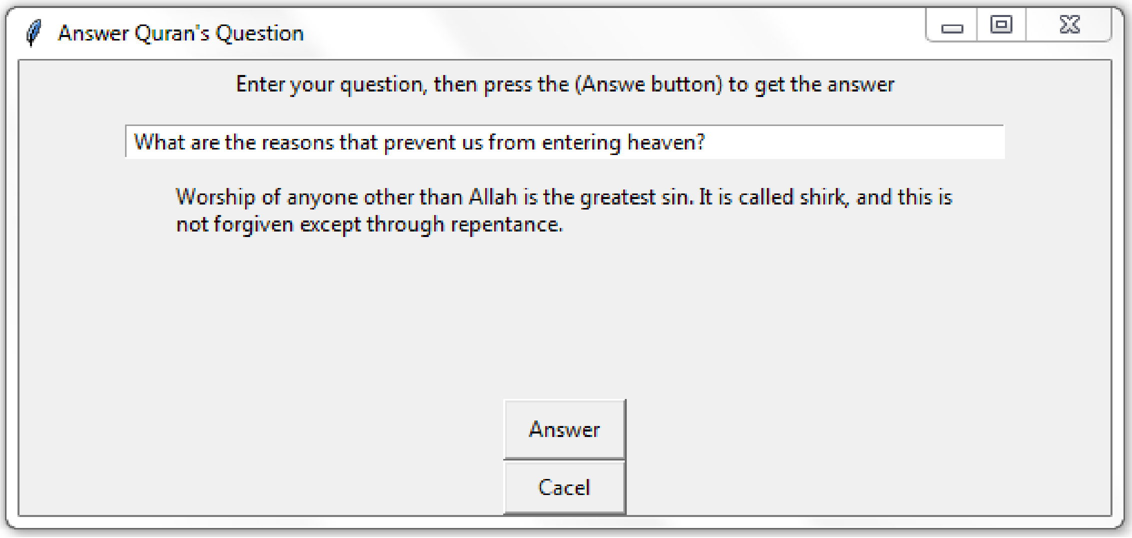


```
Python 3.4.1 (v3.4.1:c0e311e010fc, May 18 2014, 10:38:22) [MSC v.1600 32 bit (Intel
)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ================================ RESTART ================================
>>>
how will god determine who goes to heaven and hell ?
if everything is predestined, if god already knows if we are going to heaven or hel
l, why doesn't he just put us there  ?
what is it that keeps us from entering heaven?
is it really sensible that god will reward heaven only to muslims and punish others
that can easily add up to millions or even billions of people of various religions
?
what about non-muslims do, they go to heaven ?
do we believe that allah created the heavens and earth in 6 days ?
what facts about the month of ramadan have been related by hadith in regards to wha
t takes place related to heaven and hell ?
do you believe that non-muslims go to heaven ?
iam afraid of not going to heaven. i want to please god but i feel lost and confuse
d.  i was raised catholic, and i feel that if i pray the way the other religions ta
ught me god will not approve ?
goes to heaven and hell  ?  you should not feed people who are trying to find the t
rue way of islam your false ideas or your own opinion.
what do muslims think about paradise and heaven ?
if islam's teachings promote equality and being nonjudgmental, why is there a ' jud
gment ' day and some people go to heaven and some go to hell ? does the islamic rel
igion believe in the possibility of reincarnation ?
are all muslim's destined to go to heaven after they complete their time in hell ?
```

Figure 4.21: Search result after removing the stop words from the user question

Furthermore, it has been found that more relevant results were achieved after scoring and ranking the candidates. Figure4.22 shows the search result after scoring and ranking -to find the best answer- which significantly improves the search performance. It has been realized that the system produces wrong answers, when the user’s question is very short i.e. composed of few keywords, or when the context of the user question is totally different from the context of the question found in the corpus.

In the second prototype when QAEQAS has been tested using the same questions set that were used in the first version, its accuracy was more increased due to the increment of question redundancy by adding more different variants to the original question. Most questions that were got wrong answers in the first version prototype have got right answers in this version. Details have been explained in chapter 5



Enter your question, then press the (Answe button) to get the answer

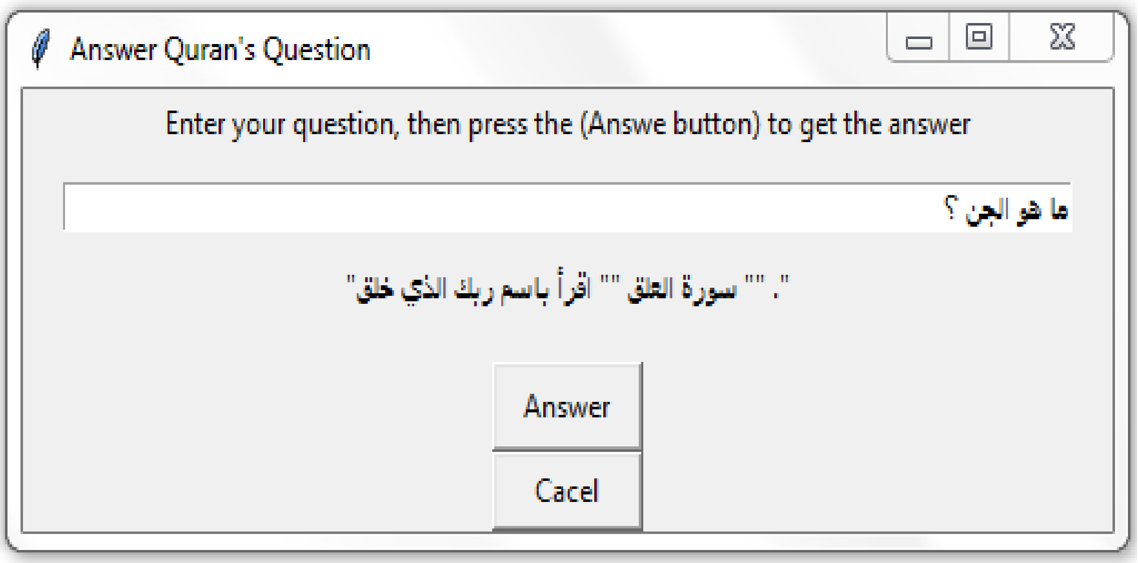
What are the reasons that prevent us from entering heaven?

Worship of anyone other than Allah is the greatest sin. It is called shirk, and this is not forgiven except through repentance.

Figure 4.22search result after scoring and ranking

We also noticed that when the user entry is not found in our corpus, instead of getting no answer, sometimes we get wrong answer because this entry may contain a word or more that are found in the corpus questions. Figure 4.23 shows the Arabic question “ما هو الجن ؟”, which is not found in our corpus. Instead of having no answer for this question, we obtained a wrong answer,”سورة العلق “which is the answer for the question: “ ما هي السورة

التي استمع لها الجن وقالوا سمعناً قرآناً عجبا. Table 4.5 shows some Arabic questions that were incorrectly answered by the system. Figure 4.24 shows a long answer given by QAEQAS.



Answer Quran's Question

Enter your question, then press the (Answe button) to get the answer

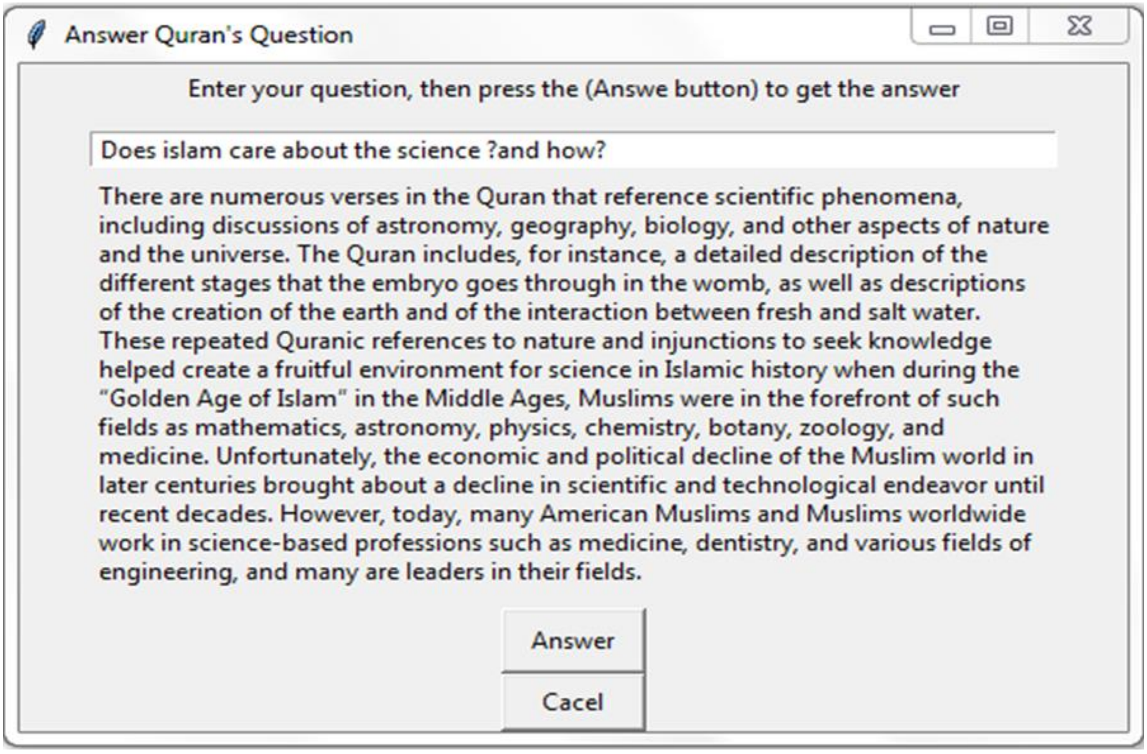
ما هو الجن ؟

"." "" سورة العلق "" اقرأ باسم ربك الذي خلق"

Figure 4.23 example of a question that has a wrong answer

Table 4.5example of Arabic questions that were incorrectly answered by the first version of the prototype

User question	Corpus question
من هو النبى الذى ذكر فى القرآن باسم اسرائيل ؟	كل الطعام كان حِلاً لبني إسرائيل إلا ما حرَّم إسرائيل على نفسه، فمن هو إسرائيل ؟
ذكر اسم صحابى من صحابة الرسول صلى الله عليه وسلم في القران الكريم فمن هو؟	من هو الصحابى الذى ذكر اسمه صراحة فى القرآن ؟
ما اقسام الانفس في القرآن	ما أنواع الأنفس التى ذكرت فى القرآن الكريم
ماهي اوهن البيوت المذكوره في القرآن	ما أضعف بيت ذكر فى القرآن الكريم



Enter your question, then press the (Answe button) to get the answer

Does islam care about the science ?and how?

There are numerous verses in the Quran that reference scientific phenomena, including discussions of astronomy, geography, biology, and other aspects of nature and the universe. The Quran includes, for instance, a detailed description of the different stages that the embryo goes through in the womb, as well as descriptions of the creation of the earth and of the interaction between fresh and salt water. These repeated Quranic references to nature and injunctions to seek knowledge helped create a fruitful environment for science in Islamic history when during the "Golden Age of Islam" in the Middle Ages, Muslims were in the forefront of such fields as mathematics, astronomy, physics, chemistry, botany, zoology, and medicine. Unfortunately, the economic and political decline of the Muslim world in later centuries brought about a decline in scientific and technological endeavor until recent decades. However, today, many American Muslims and Muslims worldwide work in science-based professions such as medicine, dentistry, and various fields of engineering, and many are leaders in their fields.

Figure 4.24 example of a question that has long answer

4.4.6 Conclusion: QAEQAS

This section presents QAEQAS,the Quranic Arabic/English Question Answering system, which takes and accepts a Natural Language (NL) question in English or Arabic from the user - through a GUI - as an input, then matches this question with the knowledge base questions, and then returns the corresponding answer. The system relies on a specialized search dataset corpus using data redundancy. Our corpus is composed of questions along with their correct answers. The questions are phrased in different ways in differing contexts to optimize the system performance. As a complete question answering solution, the Python and NLTK natural language toolkit has been used to process the user question and the corpus, as well as to implement the search engine to retrieve candidate results and then extract the best answer.

To obtain the answer, a keyword based search was used first some preprocessing techniques were applied for the user question and the corpus. Then the user question was tokenized to get the keywords, and the stop words were removed. The remaining keywords were used for searching the corpus looking for matched questions. After that, the system used scoring and ranking to find the best matched question and then return the corresponding answer for this question. QAEQAS deals with a wide range of question types including facts, definitions and it produces both short and long answers.

CHAPTER V

RESULTS ANALYSIS AND EVALUATION

5.1 Introduction

The main objectives of this research can be categorized into two targets: (i) compiling the first Islamic question and answer dataset corpus and (ii) developing an automated Arabic/English Question Answering System for the holy Quran. From the user's perspective the problem is to find the best appropriate answer for his/her question from any resource. In an ideal world, we need to measure the answers in terms of being correct and concise. As performance evaluation has been recognized as an important issue for automatic answering systems, in this chapter we mainly discuss about result analysis and performance evaluation.

Chapter 4 already includes someresults analysis and evaluation for a range of experiments in stages of our research:4.2.6 Result: QAEQ&AC; 4.2.8 Conclusions: QAEQ&AC; 4.3.1.4 Results: WEKA;4.3.1.6 Conclusions: WEKA;4.3.2.7 Conclusions: Nooj; 4.4.5 Results: QAEQAS;4.4.6 Conclusions: QAEQAS. In this chapter, we present some more detailed evaluation of our final system, QAEQAS.

In order to evaluate QAEQAS a wrong answer is not acceptable. This is because a wrong answer could provide misleading information that is not acceptable for religious reasons. This section introduces the QAEQAS evaluation method. The QAEQAS has been developed and implemented with the goal to evaluate the proposed methodology.

QAEQAS accepts English or Arabic language questions as an input. The output of QAEQAS is the answer of the best candidate question from the Quran corpus. Our methodology evaluation method uses a set of 71 questions for Arabic and 63 for English asked by user group of Muslim university academics to test both versions of the prototype against the corpus. A set of candidate questions are retrieved against each user question. When the system provides an empty set that meansit was not able to answer this question. The candidate questions are ranked according to the degree of relatedness to the user question as determined by calculating the scores. The higher the score number the higher the degree of relatedness and vice versa.

5.2 System Evaluation

5.2.1 Data Set

Restricted Domain QA System (Pre-defined Knowledge Base)

An effective way to improve the performance of QA system can be achieved by restricting the domain of questions and the size of knowledge base which resulted in the development of restricted domain knowledge base and hence restricted domain question answering system. This system overcomes the difficulties incurred in open domain and hence achieving better performance (Sasikumar, U., &Sindhu, L., 2014). As specific domain knowledge base systems are generally applied to problems that have long-term information needs for a particular domain (Dwivedi, S. K., & Singh, V. 2013), and due to the fact that the text of the Quran doesn't change; we used a pre-defined knowledge base.

Credibility of Data Set

The data set for Quran is a domain expert knowledge that composed of pairs of questions, and its correct answers from credible sources,some of these sources were mentioned in chapter 3. The answers for these questions are expert answers which have been provided by the domain experts, i.e. the Islamic Scholars. Often, scholars use references taken directly from the Quran in the form of complete or partial verses of the Quran. In most cases scholars also use multiple references to the verses to provide an answer. They usually refer to other Islamic resources such as Hadith, Sunnah and Tafsir (interpretation) of the Holy Quran. This happens because in answer to a question, the scholars usually refer to more than one verse of the Holy Quran along with references from other resources. In addition to supporting the argument, the scholars can refer to other Islamic resources such as Hadith and Tafsir in answering this question (Jilani, A., 2013).In addition to that some question along with their answers were extracted from the Quran text book as it has been explained in chapter 4.

Different Types of Question

To solve the problem concerned with answering other types of questions beyond the scope of the factoid question most of our corpus was compiled from community question answering (CQA) websites such as Islam web, TurnToIslam, ALQuran, Islamic Knowledge , etc. The CQA contain different types of questions and answers from real world. The majority of these questions can be classified as non-factoid questions.

Data Redundancy

Our approach to question answering is to take advantage of data redundancy, that already found in the corpus due to data collection from different sources, since the same question can be asked many times but in different context. In addition to that some of the corpus questions have been reformulatedin different ways in differing contexts. In the first version of the prototype each of the original questions was reformulatedso that each question is found in threedifferent variants.In the second version more data redundancy was used by reformulating each question into up to 6different variants. Tables5.1.a and 5.1.b show a question in English and Arabicrespectively that phrased in 6 different variants and has the same meaning.

Table 5.1 anExample of an English question phrased in 6different variants

Question	Answer
What happened to the previous nations when they did not obey Allah?	(29:40) So We punished each (of them) for his sins, of them were some on whom We sent Hasib (a violent wind with shower of stones) (as on the people of Lut (Lot), and of them were some who were overtaken by As-saihah (torment – awful cry. (As Thamud or Shuabs people), and of them were some when We caused the earth to swallow (as Qarun (Korah), and of them were some whom We drowned (as the people of Nuh (Noah), or Firaun (Pharaoh) and his people). It was not Allah Who wronged them, but they wronged themselves.
What is the end of the previous nations that did not obey God?	
What did God do with the folks that disobeyed Him and did many sins?	
What is the punishment that Allah has inflicted on the sinful nations?	
What is the end of the previous nations that did not obey God?	
what did Allah do for the folks who disobeyed him and make sins	

Table 5.1 b Example of an Arabic question phrased in 6 different variants

الاجابة	السؤال
زيد بن حارثة	ورد ذكر اسم صحابي من صحابة رسول الله صلى الله عليه وسلم في القران فمن هو؟
زيد بن حارثة	اذكر إسمالصحابى الذى ذكر اسمه صراحة فى القرآن الكريم؟
زيد بن حارثة	من هو الصحابي الذي ذكر اسمه في القران ؟
زيد بن حارثة	هنالك صحابي من صحابة الرسول (ص) ورد اسمه في القرآن الكريم، فمن هو؟
زيد بن حارثة	ذكر اسم صحابي من صحابة رسول الله (ص) في القرآن، من هو؟
زيد بن حارثة	من منالصحابة ورد اسمه صراحة في القرآن؟

Different results wereobtained; the accuracy of the system is increased when increasing the question redundancy. However, the greater the question redundancy in the source, the more likely it is that we can find a question in the corpus that semantically match the user question, and hence its answer. Furthermore, the larger the data set from which we can draw answers, the greater the chance we can find an answer. Given a source, that contains only one or two formulations of question may be faced with the difficult task of mapping corpus questions to user questions and hence it need to apply complex lexical, syntactic, or semantic relationships between user question string and corpus string. Therefore, we will overcome the difficulties facing when applying a complicated techniques for natural language processing. (Brill, E. et al, 2001).

The occurrence of the question in multiple different phrases (question redundancy) serves our task of question answer as follows: the occurrence of multiple linguistic formulations of the same question increases the chances of being able to find a question from different phrases that found in the corpus using a simple pattern matching the query

5.2.2 Information Retrieval

The evaluation of an information retrieval system is the process of estimating the ability of the system, how well a system satisfies the information required for the users. Many measures for evaluating the performance of information retrieval systems have been proposed. Traditional evaluation metrics, designed for Boolean retrieval or top-k retrieval, include precision and recall. Generally, measurement considers a set of documents that are searchable and the query that is being searched. Every document is known to be either relevant or irrelevant to a certain query. The selection of an appropriate retrieval model that suite the specific characteristic of the provided data is considered as a core portion of the task.

The task of the IR module is to carry out a first selection of documents (questions) that are considered relevant to the user query. Applying inadequate retrieval function would return a result where the right answer could not appear. Regular expressions in Python provide a vigorous mechanism for string matching. It is used in order to find matches among user question and Quran corpus questions. As well as the ideal ranking function should be adapted enough to fit the data characteristics (Pérez-Iglesias, et al, 2009). Ranking of query results is one of the essential problems in information retrieval.

As it has been seen before, QAEQAS displays a collection of candidate results, but we don’t know which the best ones are. One possibility is to use a similarity measure to rank the candidates. The ideal ranking method for this task should be adaptable enough to fit the specific characteristics of the data in use. One good feature is the degree of match between the user query and each of the candidates. The score for each candidate that is returned after searching was calculated according to the number of question words it contains, the more the better. Assigning scores represents the degree of results relevance in respect to the user question. Then the candidates were ranked according to their high score.

Evaluation Parameterthe system has been tested under different parameter:

Normalization

The performance of the system has been more improved after normalizing both the corpusquestions and the user question text: the noise characters were deleted; the orthography of problematic letters were unified; and short vowels and other symbols that interfere with computational manipulations with Arabic text were also handled.

Punctuation

Removing non-important punctuationis an important task in information retrieval which improved results in the form of high speed, increased relevancy, and higher accuracy and recall measures.

Tokenization

As shown in chapter 4 searching for answers in the Quran questions & answers corpus (knowledge base) without tokenizing the user’s question gives no relevant results, because the system searches for the whole question. And it gives an answer only when the user question is a full exact match to a question in the corpus, and that may happened rarely. After tokenizing the user query, many answers have been displayed, because the system searches for matches in the corpus questions for every keyword found in the user question. In this case most corpus questions will appear as candidates because these questions contain what so called stop words.

Duplicate Words

Some candidates may contain duplicate terms, which can lead to additional weight for a particular candidate. Given this situation sometimes leads to wrong answer, handling the term duplication will improve the accuracy of the system.

Stop Words

After removing standard stop words from the user question, the numbers of candidate results became less than before. When adding some common words that are found in Quran domain and the words that might be used frequently in Islamic literature,in addition to the words that used to help in building questions to the standard stop words, and then removing the entire stop words, the system displayed even fewer candidate results. This is because only few words of the user question words are remaining. These words are the important words that can be used in searching the corpus. After scoring and ranking the search result will significantly improves the search performance.

It has been realized that sometimes the system produces wrong answers;this is because: 1- the user's question is short i.e. composed of few keywords, 2- the context of the user question is totally different from the context of the question found in the corpus, this problem can be handled by increasing the question redundancy, by adding more different variants of the original question to the corpus question, 3- the same word may appear more than once in the candidate and gives it increased scores, thisproblem has been handled by counting the repeated word only once in the score. Table 5.2 shows some examples of Arabic questions that were incorrectly answered by the first version of the prototype and the reason behind that.

Table 5.2 is an example of questions that were incorrectly answered by the first version of the prototype and the reason behind that.

User question	**Corpus question**	**Discussion**
من هو النبى الذى ذكر فى القرآن باسم اسرائيل ؟	كل الطعام كان حِلاً لبني إسرائيل إلا ما حرَّم إسرائيل على نفسه، فمن هو إسرائيل ؟	The user’s question and the corpus question are different in context.
ذكر اسم صحابى من صحابة الرسول صلى الله عليه وسلم في القران الكريم فمن هو؟	من هو الصحابى الذى ذكر اسمه صراحة فى القرآن ؟	The words ” الرسول ”, ” الله ”, ” القران ” are found in many questions, and one of these questions gets higher score than the intended question.
ما اقسام الانفس في القرآن	ما أنواع الأنفس التى ذكرت فى القرآن الكريم	More than one question has the same score , the system displays the first one from the tied score
ماهي اوهن البيوت المذكوره في القرآن	ما أضعف بيت ذكر فى القرآن الكريم	There is a question in the corpus gets higher score than the intended one

Evaluation Metrics

The evaluation of an information retrieval and question answering systems is to evaluate how well a system meets the user's needs of information. Some evaluation metrics were designed for evaluating the performance of these systems. QAEQAS uses the most common evaluation metrics in Information Retrieval and Question Answering to evaluate the effectiveness and correctness of the system results, namely precision and recall. Recall typically is a measure of the percentage of relevant documents in the document set that are retrieved in response to a query, whereas precision is a measure of the percentage of retrieved documents that are relevant.

Characteristics of QAEQAS have been investigated through user-oriented testing. Evaluating QAEQAS is based on domain expert knowledge, since the corpus is composed of pairs of questions, and its correct answers from credible sources. Since we do not have a Gold Standard for the Quran questions to compare with the results, we relied on an expert user group of Muslim university academics to ask QAEQAS some questions in our questions domain. Table 5.3 details the evaluation of the two versions of the prototype, for Arabic and English questions with the answers retrieved by the system; Table 5.4 shows the accuracy of the two versions of QAEQAS prototype.

Table 5.3 Evaluation of the two versions of QAEQAS prototype

Question Redundancy / No. of Questions	The first version, the question in 3 different variants		The second version, the question in 6 different variants	
	Arabic	English	Arabic	English
Total number of questions	71	63	71	63
number of correctly answered questions	54	46	67	56
number of incorrectly answered questions	14	15	3	6
number of answers the system failing to return	3	2	1	1
total number returned answer	68	61	70	62

Accuracy for the First Version of the Prototype

Arabic

Precision = rlv / ra → 54/68 = 79%

Recall = rlv/ ta → 54/71 = 76%

English

Precision = rlv / ra → 46/61 = 75%

Recall = rlv/ ta → 46/63 = 73%

Accuracy for the Second Version of the Prototype

Arabic

Precision = rlv / ra → 67/70 = 96%

Recall = rlv/ ta → 67/71 = 94%

English

Precision = rlv / ra → 56/62 = 90%

Recall = rlv/ ta → 56/63 = 89%

rlv = the number of correctly answered questions (relevant)

urlv = the number of incorrectly answered questions (irrelevant)

ra = the total number returned answer (rlv+urlv)

urta = the number of answers the system failing to return (unreturned)

ta = the total number of questions (returned + unreturned)

Table 5.4 the accuracy of the two versions of QAEQAS prototype

Accuracy by language **Redundancy**	**Arabic**		**English**	
	Precision	**Recall**	**Precision**	**Recall**
3 variants of the question	79%	76%	76%	73%
6 variants of the question	96%	94%	90%	89%

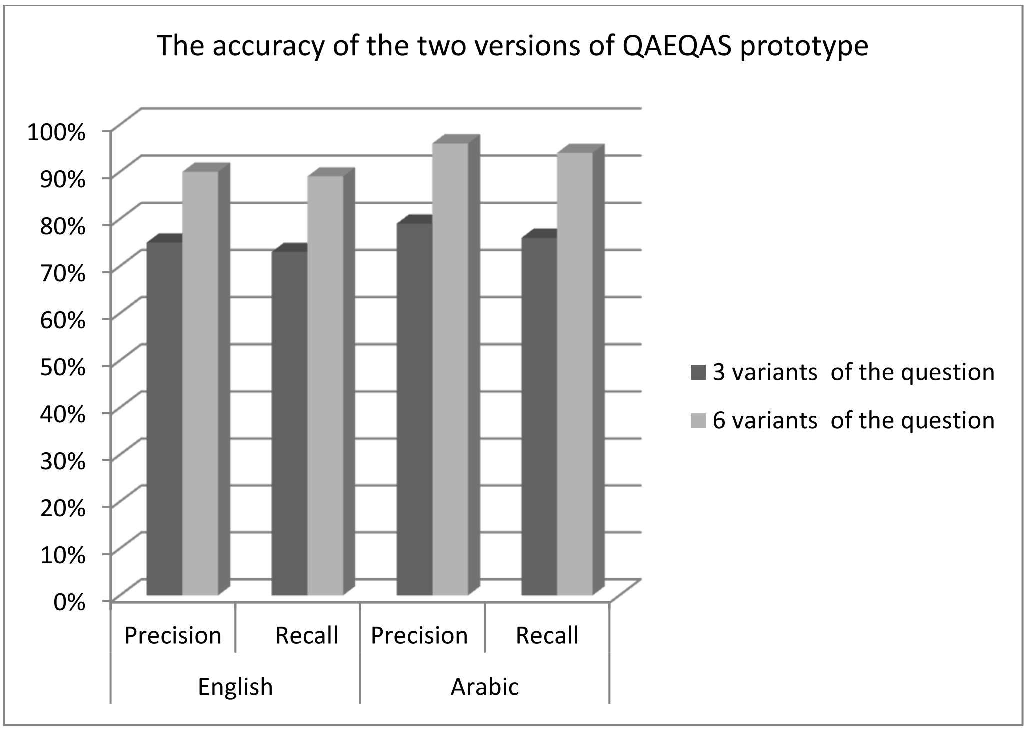


Figure 5.1 the accuracy of the 2 versions of QAEQAS prototype

From Table 5.4 we can see that the accuracy of the systemis increased by increasing the data redundancy, Table 5.5 shows the increment in precision and recall due to data redundancy for both Arabic and English.

Table 5.5 the increment in precision and recall due to data redundancy

Arabic		English	
Precision	Recall	Precision	Recall
17%	18%	14%	16%

From the above results,we can realize that both precision and recall are higher for the Arabic Language than it is for English, although the processing of the Arabic language is more difficult than the English language. Arabic is the original language of the Quran, and religious scholars may suggest this is a reason why Quran QA in Arabic works better, because Arabic is the desired language for this message. We can speculate or guess at reasons but we cannot prove readily what the interpretation should be of these results:

- Problems of Arabic characters such as the orthography of problematic letters, the short vowels and other symbols (harakat) that interfere with computational manipulations has been addressed.
- The Arabic stop words which created byTahaZerrouki are too much than the English stop words list that included within NLTK.
- In addition to that the Arabic stop words are containing all forms of stop words.

5.3Conclusions: Evaluation of QAEQAS

This chapter reviews the analysis of the results obtained from the experiments that has been conducted in the previous chapter to solve the research problem. Furthermore it presents the evaluation of the proposed solution, and shows its usefulness in dealing with a wide range of question types, with a precision of 96% and a recall of 94 for Arabic; and a precision of 90% and a recall of 89% for English.

CHAPTER VI

CONCLUSION AND FUTURE WORK

6.1 Overview

6.1.1 Overall Findings

Muslims believe that the Quran is the most authoritative source for knowledge, guidance, rule, sole and legislations for all human. It has been a big challenge for the computer scientists to design accurate systems that answer users question in any domains. In this thesis we developed QAEQAS a question answering system that can answer the user question about the Quran in both Arabic and English languages. First we have developed a special data set of around 1500 questions along with their correct answers compiled from scholarly sources. This data set has been used later in QAEQAS, and can be used in text mining applications where Quran questions and answers have to be searched.

Different alternative NLP technologies has been examined and explored to investigate how suitable they are to build a QA system for answering questions about the Quran. Some clustering work has been done using WEKA tool, in which the relevant attributes in the Q&A data set were selected to decide the cluster. Furthermore aNoojmodel has been developed to presents some solutions for automating the processing of an English version of Quran Q&A corpus. Moreover, Python and natural language toolkit (nltk) has been used to automate the processing of the user question and the corpus; as well as to implement the search engine to retrieve candidates result and then extract the best answer.This new resource and models presented yet for Quran question answering system will set up a reliable foundation for many interesting linguistic and corpus-based studies.

QAEQASanswers all questiontypes and returns an acceptable short and long answers.From previous results, we can realize that the system performance is increased by increasing data redundancy: both precision and recall are high for Arabic and English. But the results of QAEQAS are different for English and Arabic this is because the processing steps are different for each.QAEQAS presents its usefulness: efficient and effectiveness that deals with a wide range of question types in addition to its speed and high accuracy.

Thisresearch focuses on building a resource that would enter into beneficial text mining applications. The main contribution of this thesis has been the development of novel resource, namely, a Quranic Arabic-English Question and Answer Corpus (QAEQ&AC)that compiled from scholarly sources. This corpus was used as a knowledge base to developQAEQAS (Quranic Arabic English Question Answering System). This corpus can alsobe used in text mining applications and machine learning where Quran questions and answers are needed.

6.2 Future Work

I am looking forward to continue the work on the topic explored in this thesis;we spenta lot of time and effort into the task of compiling the Quran Q&A dataset: collecting questions along with accurate answers, merging reformatting, cleaning, and converting it to required format. After having these dataset at my disposal a number of experiment and text processing task were conducted. we left out a number of possible tasks as future improvements. The following provide an outline.

6.2.1 Improvements on Quran Q&A Dataset

The work in this thesis covers the questions and answers of the Quran as a whole, which makes it a huge domain for further work. This scope could be not cover all questions and answers of the Quran, later can be enhanced to cover the whole questions and answers of the Quran. The Quran Q&A corpus was developed as a first and second version, in future this dataset can be further improved in a number of ways:

Extending the Quranic Q&ACorpus

- The Arabic and English Corpus can be extended by adding more questions and answersfrom credible sources
- Increase the question redundancy by adding more phrases in many different ways in differing contexts to all corpus questions in order to further improve the performance of the system.
- Compiling more Q&A dataset corpus in other languages

Validation

Most of questions and answers of this dataset were created from scholars works found on the internet and hence one would assume they were accurate and verified. However, they need to be subjected to further manual validation by Quranic scholars before making it available on the internet and incorporating them into wider applications.

6.2.2 Collecting Extra QuranicCorpus

In addition to our Q&A corpus, Quran QA system require more pre-constructed knowledge sources, such as domain ontology, Quranic corpus, which combines the Holy Quran, Hadith, and other parts of the Islamic faith.

6.2.3 Understanding the User Question

The Quran may use many different words to describe the same concept, or one term to point to completely different meanings, depending on the context of the sentence. Hence complex linguistic analysis of user question can be applied in order to understand as much as possible about the meaning of the questions being asked.

6.2.4 Extending QA to Other Religious Texts

The system can be extended to cover other religious texts such as Bible, by collecting question and answer,and also other text corpus for Bible or other religious.

REFERENCES

Abbas, N.H., 2009. *Quran'search for a concept'tool and website* (Doctoral dissertation, University of Leeds (School of Computing)).

Abdelnasser, H., Mohamed, R., Ragab, M., Mohamed, A., Farouk, B., El-Makky, N. and Torki, M., 2014. Al-Bayan: an arabic question answering system for the holy quran. *ANLP 2014*, p.57.

Aggarwal, C.C. and Zhai, C. eds., 2012. *Mining text data*. Springer Science & Business Media.

Alfaifi, A. and Atwell, E., 2014. Tools for Searching and Analysing Arabic Corpora: An evaluation study. *Learning and Teaching for Right to Left Scripted Languages: Realities and possibilities*, p.21.

AlMaayah, M., Sawalha, M., &Abushariah, M. (2014). A proposed model for Quranic Arabic WordNet. In *Proceedings of the 2nd Workshop on Language Resources and Evaluation for Religious Texts, 31 May 2014, Reykjavik, Iceland*(pp. 9-13). LRA.

All Quran "Islamic material/frequently asked questions(FAQ)", http://www.all-quran.com/islamic_material/frequently_asked_questions.html,30/5/2015

Anita Wasilewska, Jae Hong Kil (105228510),

Atwell, E., Brierley, C., Dukes, K., Sawalha, M. and Sharaf, A.B., 2011. An Artificial Intelligence approach to Arabic and Islamic content on the internet. In *Proceedings of NITS 3rd National Information Technology Symposium* (pp. 1-8). Leeds.

Atwell, E.S., Brierley, C. and Sawalha, M., 2012. Proceedings of LREC'2012 Workshop LRE-Rel: Language Resources and Evaluation for Religious Texts.

Banko, M., Brill, E., Dumais, S. and Lin, J., 2002, March. Askmsr: Question answering using the worldwide web. In *Proceedings of 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases* (pp. 7-9).

Bogdanova, D., Ganguly, D., Foster, J., &Vahid, A. H. (2015). *ADAPT. DCU at TREC LiveQA: A Sentence Retrieval based Approach to Live Question Answering*. Dublin City University Dublin Ireland.

Bird, S., Klein, E. and Loper, E., 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Bird, S., Klein, E., &Loper, E.Preface".www.nltk.org. Updated for nltk 3.0. This is a chapter from Natural Language Processing with Python. Copyright © 2014 the authors. It is distributed with the Natural Language Toolkit [http://nltk.org/], Version 3.0. retrieved 2016-4-15

Blooma, M.J. and Kurian, J.C., 2012. Clustering Similar Questions In Social Question Answering Services. In *PACIS* (p. 160).

Bouckaert, R.R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A. and Scuse, D., 2013. Waikato Environment for Knowledge Analysis (WEKA) Manual for Version 3-7-8. *The University of Waikato, Hamilton, New Zealand*.

Brill, E., Lin, J.J., Banko, M., Dumais, S.T. and Ng, A.Y., 2001, November. Data-Intensive Question Answering. In *TREC* (Vol. 56, p. 90).

Brill, E., Dumais, S., &Banko, M. (2002, July). An analysis of the AskMSR question-answering system. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 257-264). Association for Computational Linguistics.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., 2000. CRISP-DM 1.0 Step-by-step data mining guide.

Comma Separated Values (CSV) Standard File Formathttp://edoceo.com/utilitas/csv-file-format , time of access 14/6/2015

CSV file format Csvreader.com, http://www.csvreader.com/csv_format.php., time of access 17/6/2015

Data Mining http://www.unc.edu/~xluan/258/datamining.html

Data Mining http://www.unc.edu/~xluan/258/datamining.html, 17/6/2015

Dominic John Repici, 2010.The Comma Separated Value (CSV) File Format. http://creativyst.com/Doc/Articles/CSV/CSV01.htm

Dukes, K. and Atwell, E., 2012. LAMP: a multimodal web platform for collaborative linguistic analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (pp. 3268-3275). European Language Resources Association (ELRA).

Dukes, K., Atwell, E. and Sharaf, A.B.M., 2010, May. Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. In *LREC*.

Dwivedi, S.K. and Singh, V., 2013. Research and reviews in question answering system. *Procedia Technology*, *10*, pp.417-424.

Gabrilovich, E., &Markovitch, S. (2007, January). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI* (Vol. 7, pp. 1606-1611).

Graesser, A. C., Rus, V., Cai, Z., & Hu, X. (2011). Question answering and generation. Applied NLP. IGI Global, Hershey.

Gusmita, R.H., Durachman, Y., Harun, S., Firmansyah, A.F., Sukmana, H.T. and Suhaimi, A., 2014, November. A rule-based question answering system on relevant documents of Indonesian Quran Translation. In *Cyber and IT Service Management (CITSM), 2014 International Conference on* (pp. 104-107). IEEE.

Hamoud, B. and Atwell, E., 2016, February. Quran question and answer corpus for data mining with WEKA. In *Basic Sciences and Engineering Studies (SGCAC), 2016 Conference of* (pp. 211-216). IEEE.

Hamoud, B. and Atwell, E.S., 2016, March. Compiling a Quran Question and Answer Corpus: تجميع مدونة اسئلة واجوبة للقرآن الكريم. In *ICCA'2016 International Conference on Computing in Arabic*. Leeds

Hamoud, B. and Atwell, E.S., 2016. Using an Islamic Question and Answer Knowledge Base to answer questions about the holy Quran. *International Journal on Islamic Applications in Computer Science And Technology*.

Hamoud, B. and Atwell,E.S., 2017. Evaluation corpus for restricted-domain question-answering systems for the holy Quran, International Journal of Science and Research (IJSR), Volume 6 Issue 8, August 2017

http://www.creativyst.com/Doc/Articles/CSV/CSV01.htm

http://www.sultan.org/." Discover Islam and Muslim Beliefs. Learn about The Real Islam. Correct your information about Islam Religion"

Islamic Knowledge/Come towards Islam "questions and answers about Quran", https://islamicknowledge2all.wordpress.com/2011/10/30/question-and-answers-about- quran-3/, time of access 30/5/2015

Islamic question and answer (http://islamicquestions.net/)

Jagtap, S.B., 2013. Census data mining and data analysis using WEKA. *arXiv preprint arXiv:1310.4647*.

Jilani, A., 2013. *Parallel corpus multi stream question answering with applications to the Qu'ran* (Doctoral dissertation, University of Huddersfield).

Kanaan, G., Hammouri, A., Al-Shalabi, R. and Swalha, M., 2009. A new question answering system for the Arabic language. *American Journal of Applied Sciences*, *6*(4), p.797.

Khan, H.U., Saqlain, S.M., Shoaib, M. and Sher, M., 2013. Ontology based semantic search in Holy Quran. *International Journal of Future Computer and Communication*, *2*(6), p.570.

Kangavari, M. R., Ghandchi, S., &Golpour, M. (2008). A new model for question answering systems. *World Academy of Science*, *Engineering and Technology*, *42*, 506-513.

Leskovec, J., Rajaraman, A. and Ullman, J.D., 2014. *Mining of massive datasets*. Cambridge university press.

Li, S., 2011. *Beyond Question Answering: Understanding the Information Need of the User* (Doctoral dissertation, University of York).

Lundh, F. (1999). An introduction to tkinter. *URL: www. pyhonware. com/library/tkinter/introduction/index. htm*.

Manning, C.D., Raghavan, P. and Schutze, H., 2009. An information to information retrieval.

Miner, G., 2012. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press

Muhammad, A.B., 2012. *Annotation of conceptual co-reference and text mining the Qur'an*. University of Leeds.

Pavan, M. and Todeschini, R., 2008. *Scientific data ranking methods: theory and applications* (Vol. 27). Elsevier..

Pérez-Iglesias, J., Garrido, G., Rodrigo, Á., Araujo, L. and Peñas, A., 2010. Information retrieval baselines for the ResPubliQA task. *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pp.253-256..

Perkins, J., 2014. *Python 3 Text Processing with NLTK 3 Cookbook*. Packt Publishing Ltd..

Pujar, S., Priyaa, B. and Sethia, K., Distributed QA System.

question and answer about Islam. Islam Question and answer. http://islamqa.info/en/

Question Answering. https://en.wikipedia.org/wiki/Question_answering. Time of access 12/01/2015

Ravichandran, D., Ittycheriah, A., &Roukos, S. 2003. Automatic derivation of surface text patterns for a maximum entropy based question answering system. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2* (pp. 85-87). Association for Computational Linguistics.

Raza, K., 2012. Application of data mining in bioinformatics. *arXiv preprint arXiv:1205.1125*.

Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald.”WEKA Manual for Version 3-7-11” university of WAIKATO

Roberts, A., Al-Sulaiti, L. and Atwell, E., 2006. aConCorde: Towards an open-source, extendable concordancer for Arabic. *Corpora*, *1*(1), pp.39-60.

RouseMargaret, 2017 Natural language processing (NLP), http://searchcontentmanagement.techtarget.com/definition/natural-language-processing-NLP,

Saad, M.K. and Abed, R.M., 2012. Distributed data mining on grid environment. *American Academic & Scholarly Research Journal*, *4*(5), p.1.

Sabbah, T. and Selamat, A., 2014, September. Support vector machine based approach for Quranic words detection in online textual content.

In *Software Engineering Conference (MySEC), 2014 8th Malaysian* (pp. 325-330). IEEE.

Sapp, B., Saxena, A. and Ng, A.Y., 2008, July. A Fast Data Collection and Augmentation Procedure for Object Recognition. In *AAAI* (pp. 1402-1408).

Sasikumar, U. and Sindhu, L., 2014. A Survey of Natural Language Question Answering System. *International Journal of Computer Applications*, *108*(15).

Sharaf, A.B.M. and Atwell, E., 2012. QurAna: Corpus of the Quran annotated with Pronominal Anaphora. In *LREC* (pp. 130-137).

Sharaf, A.B.M. and Atwell, E., 2012. QurSim: A corpus for evaluation of relatedness in short texts. In *LREC* (pp. 2295-2302).

Shawar, B.A. and Atwell, E., 2009. Arabic question-answering via instance based learning from an FAQ corpus. In *Proceedings of the CL 2009 International Conference on Corpus Linguistics. UCREL* (Vol. 386).

Sherkat, E. and Farhoodi, M., 2014, October. A hybrid approach for question classification in Persian automatic question answering systems. In *Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on* (pp. 279-284). IEEE.

Shmeisania, H., Tartirb, S., Al-Na'ssaanc, A. and Najid, M., Semantically Answering Questions from the Holy Quran.

Song, W., Feng, M., Gu, N. and Wenyin, L., 2007, October. . In *Semantics, Knowledge and Grid, Third International Conference on* (pp. 298-301). IEEE.

Sulistyanto, H. and Azhari, S.N., 2013. A Few Survey of Developments and Challenges Arising on General and Indonesian Question Answering System.

The National Centre for Text Mining (NaCTeM) at The University of Manchester, 2014. Text mining tools. http://argo.nactem.ac.uk

The siasat daily "Questions and answers about the holy Quran", http://www.siasat.com/english/news/questions-answers-about-holy-quran?page=0%2C0,30/5/2015

Turn to Islam Community, "questions-on-Quran", http://turntoislam.com/community/threads/100-questions-on-quran.10052,time of access 30/5/2015

Ullman, J. D., Leskovec, J., &Rajaraman, A. (2011). Mining of Massive Datasets

Wikipedia, "Comma separated value", https://en.wikipedia.org/wiki/Comma-separated_values

Voorhees, E. M., & Tice, D. M. 1999, The TREC-8 Question Answering Track Evaluation. In *TREC* (Vol. 1999, p. 82).

Witten, I. H., & Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Witten, I.H., Frank, E., Hall, M.A.” Data mining practical machine learning tools and techniques, 2011” (third edition) Morgan Kaufmann .

Yih, W. T., & Ma, H. 2016. Question Answering with Knowledge Base, Web and Beyond. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 1219-1221). ACM.

Yunus, M.A., Zainuddin, R. and Abdullah, N., 2010, December. Semantic query for Quran documents results. In *Open Systems (ICOS), 2010 IEEE Conference on* (pp. 1-5). IEEE..

Zhenqiu, L., 2012. Design of automatic question answering system base on CBR. *Procedia Engineering*, *29*, pp.981-985.

تجميع مدونة اسئلة واجوبة للقرآن الكريمAppendix A

المستخلص_في هذهالورقة وصفنا تجميع مجاميع (corpus) الأسئلة والأجوبة للقرآن الكريم من خلال دمج مجموعات بيانات (datasets) مختلفة. قمنا بتجميع الأسئلة مع إجاباتها باللغتين العربية و الإنجليزية وذلك من عدة مصادر، ثم دمجها و تنظيفها وتحويلها الى التنسيق الذي يناسب أدوات تحليل النص، وذلك باستخدام ميكروسوفت اكسل. قمنا أولا بإدخال كافة البيانات الخام لكل لغة على حدا بورقة عمل منفصلة بملف اكسل، ثم تنظيفها وذلك بإزالة البيانات غير ذات الصلة وغير المتناسقة وكذلك البيانات التي لا نريدها. استخدمنا في ذلك طريقتين: الطريقة اليدوية، والطريقة الآلية على سبيل المثال لإزالة البيانات غير الصحيحة أو غير الكاملة استخدمنا الطريقة اليدوية حيث ان استخدام الطرق الآلية في مثل هذه الحالة قد يكون غير مجديا. لإزالة المسافات الزائدة وفواصل الأسطر،أو حذف الأحرف غير القابلة للطباعة الخ. استخدمنا الطرق الآلية حيث ان الطرق اليدوية قد تستغرق ساعات مضنية أو قد لا تضمن اكتشاف و إزالة كل الأخطاء. ثم بعد ذلك تم تحويل البيانات الى التنسيق الذى يناسب أدوات التعدين، حيث قمنا بإنشاء ملف بتنسيق قيمة مفصولة بفواصل (Comma Separated Value(CSV)). لإعداد المجاميع (corpus) اتبعنا نهج تدريجي وذلك بالاحتفاظ بالبيانات بعد كل مرحلة من مراحل إعدادها حتى توصلنا الى بيانات جاهزة للتحليل دون تدمير مجموعة البيانات الأصلية. بهذا الاسلوب يمكننا الحفاظ بتوثيق كامل للبيانات حتى يتثنى الرجوع اليها في وقت لاحق أو من قبل أشخاص آخرين. بعد هذه المرحلة يمكن تطبيق تقنيات تجهيز البيانات (data preprocessing)باستخدام أدوات تنقيب البيانات أو أدوات تحليل النص و ذلك استعدادا لمزيد من التحليل.

1المقدمة

القرآن هو كلام الله المنَّزل بواسطة الملاك جبريل على رسول الإسلام وخاتم النبيين محمد (ص) بلغة العرب. وهو الكتاب الذي يؤمن به المسلمون، فهو محفوظ من الله من كل مس أو تحريف وهو آخر الكتب السماوية ويعد القرآن أرقى الكتب العربية قيمة لغوية ودينية، لما يجمعه من البلاغة والبيان والفصاحة. يحتوي القرآن على رسالة الله للبشرية جمعاء حيث أنه يخاطب الأجيال كافة و في كل القرون. فهو المصدر الرئيسي للتوجيه واللوائح والقوانين حيث أنه يتضمن كل المناسبات ويحيط بكل الأحوال. فالقرآن ينص على أحكام العقائد، وفيه أحكام العبادات و المعاملات والأخلاق والآداب.

لا يقتصر تفسير القرآن على الجانب الحرفي بل يمتلك المظهر الخارجي وعمق مخفي. في بعض الأحيان يصعب على المستخدم فهم بعض المعاني ولا يستطيع أن يجد جوابه مباشرة من القرآن ، بل عليه أن يتحمل عبء البحث في عدد من الكتب الإسلامية مثل الحديث، التفسير وغيرهما للعثور على الإجابة المطلوبة، مما يجعل البحث مملا ومهدرا للوقت. يفضل الكثير من المستخدمين استخدام الإنترنت للإجابة على أسئلتهم فيكون البحث في مجموعة ضخمة من البيانات تتكون من العديد من الكتب الإسلامية الإلكترونية أو مواقع الإنترنت. بعد كل هذا البحث قد لا يجد المستخدم ما يصبو إليه من الإجابة. لم يكن هنالك ما يكفي من الموارد الحالية ما هو مصمما خصيصا لأسئلة وأجوبة القرآن الكريم. وحتى هذه الموارد الموجودة تتناثر بين صفحات الويب المختلفة. كل من هذه المصادر له شكله الخاص واسلوبه. هنالك حاجة إلى إنشاء مجاميع مجموعة بيانات موحدة لأسئلة وأجوبة القرآن ليتثنى سهولة استخدامها في الاختبار والتقييم في تطبيقات ابحاث القرآن أو أنظمة الرد على أسئلة القرآن.

المجاميع (corpus) عبارة عن مجموعة كبيرة من النصوصالمنظمة، والتي يمكن ان تستخدم للقيام بالتحليل الإحصائي، واختبار الفرضيات، والتحقق من الوقائع أو التحقق من صحة قواعد لغوية داخل لغة معينة، ويمكن ان يكتب نص المجاميع بلغة واحدة أو أكثر. تستخدم أدوات تحليل المجاميع على نطاق واسع من قبل الباحثين لدراسة النص الأصلي والنص الذي تمت ترجمته ، فإنها تسمح للمستخدمين بالوصول إلى المعلومات الواردة ضمن المجاميع وعرضها ودراستها بعدة طرق. غالبا ما تتكون حزم تحليل المجاميع (corpus) من مجموعة متنوعة من الأدوات مثل نووج، آرقو، ويكا، الخ.

2جمع البيانات

الخطوة الأولى لبناء المجاميع (corpus) هو التفكير في مصادر البيانات وكيفية جمع هذه البيانات. عملية جمع البيانات يمكن أن تكون بسيطة نسبيا وفقا لنوع الأدوات المستخدمة في جمعها. هناك العديد من الأدوات التي يمكن استخدامها لجمع البيانات، يجب أن تكون هذه الأدوات جيدة بما فيه الكفاية لجمع بيانات مفيدة من أجل الحصول على تقييم أفضل للبحوث. اختيار أدوات محددة، تعتمد على طبيعة المهمة، فضلا عن نوع البيانات المطلوبة. في هذه الورقة تم استخدام أربعة مصادر لجمع أسئلة و أجوبة من القرآن الكريم، و حيث ان الإنترنت غنيا بالبيانات مع امكانية الوصول إليها بسهولة، كان المصدر الأول والرئيسي لجمع المجاميع (corpus) لدينا هو مواقع على شبكة الإنترنت تم إنشاؤها من قبل مجموعة من الباحثين في المجال الإسلامي. المصدر الثاني هو مجموعة من المسجد الحرام في مكة المكرمة. والثالث هو من مسح البحوث السابقة لأنظمة أسئلة واجوبة من القرآن. جمع البيانات هو جزء لا يتجزأ للعديد من المشاريع البحثية، فوجود بيانات كافية للتعلم ذي أهمية كبيرة على الأداء الجيد وتقييم نجاح أو

فشل النظام. كمية ونوعية بيانات التدريب من الاهمية بمكان لتعلم الخوارزميات، فزيادة حجم مجموعة التدريب وتطوير خوارزميات تعلم أفضل له مساهمة كبيرة في تحسين الأداء [1].

البحث في المجال الديني للرد على الأسئلة ينطوي على المخاوف الأخلاقية. الإجابة على أسئلة حول المعتقدات الإسلامية يتطلب عناية فائقة لإعطاء إجابة دقيقة لسؤال معين، حتى تكون مقبولة دينيا . على سبيل المثال ينبغي أن تذكر الإجابات على النحو المذكور في القرآن و كتب الحديث. جمع أسئلة واجوبة من مصادر موثوقة وذات مصداقية مسألة هامة حيث ان الدقة المنخفضة أو الإجابة الخاطئة ليست مقبولة في المجال الديني وخاصة في مجال القرآن الكريم. في هذه الورقة استخدمنا الأسئلة التي تتكرر كثيرا (FAQ) من المصادر المذكورة أعلاه من أجل جمع الأسئلة والأجوبة. لقد دمجنا مجموعات فرعية مختلفة من البيانات من مصادر مختلفة لتشكيل مجموعة بيانات موحدة للقرآن. بدأنا بالبحث عن الموارد على شبكة الإنترنت، واخترنا ما يلي:

•turntoislam [2] وهو موقع شعبي على شبكة الإنترنت لمعرفة المزيد عن الإسلام، يحتوي على مكتبة ضخمة والتي تتكون من العديد من المواضيع حول الإسلام ترجمت الى العديد من اللغات. وهو يتضمن أجوبة على العديد من الأسئلة، وأشرطة الفيديو، واستطلاعات الرأي، والأحداث، الى جانب بعض الاشياء الاخرى. ويهدف إلى إظهار جمال الإسلام للعالم، فضلا عن بناء مجتمع ودي، مطبقا للقيم الإسلامية.

• Islamic Knowledge/Come towards Islam[3]وهو موقععلى شبكة الإنترنت و يحتوي على أرشيف شهري يغطي العديد من المواضيع الاسلامية مثل أسئلة وأجوبة حول القرآن، وفهم الإسلام، والحقائق الإسلامية، ومناقشة سور القرآن، تعاليم النبي محمد (ص) ، المحرم من المواد

الغذائية والمشروبات، شهر رمضان، المرأة في الإسلام وغيرها. له ارشيف من مارس 2011 حتى فبراير 2015. وهذا الموقع يوفر أيضا ترجمة لنص القرآن في العديد من اللغات، كوسيلة للقراءة والاستماع للقرآن على الإنترنت.

- All–Quran [4] ويهدف الموقع إلى أن يكون القرآن الكريم متاحا للجميع من خلال توفير وسيلة سهلة لتدفق الصوت لمجموعة متنوعة من قراء القرآن والترجمات الصوتية. أنه يحتوي على علامة تبويب التعليمات الإسلامية، بالإضافة إلى النص المشروح من القرآن الكريم. وهو موقع مجانيا على شبكة الإنترنت.

- Siasat daily[5] سياسات اليومية هي موقع على شبكة الإنترنت يوفر أسئلة وأجوبة حول القرآن الكريم. وهو مكتوب بثلاث لغات، الإنجليزية، الأردية، والهندية.

- SULTAN ISLAMIC LINKS[6] تحوي العديد من المواضيع على سبيل المثال، اكتشف الإسلام، والشعب المسلم، القرآن الكريم والأديان الإسلامية. لديها الرابط "أنت تسأل والقرآن يجيب"

- Islamic question and answer [7] أسئلة وأجوبة عن الإسلام فهي تحتوي على العديد من الأسئلة حول الإسلام ترجمت الى 13 لغة.

نحن لا نعتمد حصرا على مصادر شبكة الإنترنت، بل جمعنا مجموعة من الأسئلة التي قام بإجابتها عدد من الشيوخ في المسجد الحرام بمكة المكرمة، والذين هم قادة في مجال الدراسات الإسلامية. تعطي هذه المجموعة إجابات خبيرة على الأسئلة التي تطرح بواسطة المسلمين الذين يأتون إلى المسجد الحرام. الى جانب ذلك أدرجنا مجموعات اختبار صغيرة من الأسئلة والأجوبة من الأبحاث السابقة.

3إعداد البيانات

غالبا ما تحتوي البيانات على العديد من الأخطاء مما يؤدي إلى استنتاجات خاطئة عند تحليلها، وذلك بعد اهدار الكثير من الوقت، حيث ان الأخطاء قد تكون في البيانات نفسها وليست في التحليل . قد يحتاج تنظيف البيانات الى نصف الوقت اللازم لتحليلها. و بمجرد الحصول على بيانات نظيفة، سوف يتم التحليل بصورة سلسة وواضحة تماما. تنظيف البيانات وإعدادها للتحليل هي واحدة من تلك الوظائف الصعبة فهي مملة، شاقة، ومضنية، بغض النظر عن الطريقة التي تستخدم، حيث ان تكلفة الخطأ معتبرة جدا. ونحن نعتقد أن استخدام ميكروسوفت اكسل من اسهل الطرق لتنظيف البيانات حيث يمكننا اتباع نهج تدريجي وذلك بالاحتفاظ بالبيانات بعد كل مرحلة من مراحل إعدادها بورقة عمل منفصلة أو بملف منفصل حتى نصل الى بيانات جاهزة للتحليل دون تدمير مجموعة البيانات الأصلية. بهذا الاسلوب يمكننا الاحتفاظ بتوثيق كامل للبيانات حتى يتثنى الرجوع اليها في وقت لاحق أو من قبل أشخاص آخرين.

لا توجد حاليا موارد مصممة خصيصا لأسئلة واجوبة القرآن الكريم فقد قمنا بدمج مجموعات فرعية مختلفة من البيانات من اجل انشاء مجموعة مشتركة، تتألف هذه المجموعة من البيانات الجديدة وبعض البيانات الموجودة من مختلف مجموعات اختبار صغيرة. إعداد البيانات يغطي جميع المهام لبناء مجموعة البيانات النهائية من البيانات الخام الأوليةمثل الجداول والسجلات واختيار سمات (instances) محددة، بالإضافة الى تحويل أو بناء، وتكامل البيانات ، وتنظيفها وإعادة تهيئتها لأدوات النمذجة. مهام إعداد البيانات يمكن أن يؤدى عدة مرات، وليس في ترتيب معين [8] فيما يلي مراحل إعداد البيانات التي قمنا بها:

3.1 إنشاء ملف البيانات

مجموعات البيانات(data sets) التي تم جمعها لها اشكال وتنسيقات مختلفة، فالخطوة التالية هو دمجها و توحيدها باستخدام ميكروسوفت اكسل 2010. لقد قمنا بإدخال كافة مجموعات البيانات الاصلية (الخام) بورقة عمل من مصنف اكسل ولكل لغة على حدا، لقد استخدمنا في ذلك اللصق أوالاستيراد لبعضها وكتابة البعض الآخر يدويا.

3.2 تنظيف البيانات

تنظيف البيانات، وتسمى أيضا تطهير البيانات أو غسل البيانات، وهى عملية كشف و إزالة الأخطاء والتناقضات من البيانات من أجل تحسين نوعية البيانات. مشاكل جودة البيانات موجودة في مجموعات بيانات مفردة، مثل الملفات وقواعد البيانات، وذلك بسبب أخطاء إملائية أثناء إدخال البيانات، معلومات ناقصة أو بيانات غير صالحة. اما بالنسبة للبيانات التي يتم دمجها من عدة مصادر مثل مستودعات البيانات، أو نظم المعلومات على شبكة الإنترنت العالمية تزيد الحاجة الى تنقية البيانات بشكل كبير ويعزى ذلك لوجود مصادر بيانات غير متجانسة والتي غالبا ما تحتوي على بيانات مكررة أو زائدة عن الحاجة وفي تمثيلات مختلفة.

تنظيف البيانات هي المرحلة التي يتم فيها إزالة البيانات التي ليس لها صلة، و الضجيج من مجموعة البيانات، فهي عبارة عن اجزاء من البيانات غير صحيحة أو ناقصة أو غير دقيقة الخ ، ومن ثم استبدال أو تعديل أو حذف هذه البيانات. تنظيف البيانات هي عملية كشف وتصحيح (أو إزالة) سجلات فاسدة أو غير دقيقة من مجموعة سجلات أو جدول أو قاعدة بيانات. قد تحتوي البيانات غير النظيفة على الأخطاء، مثل الأخطاء الإملائية أو أخطاء في العلامات

مثل الفاصلة والنقطة والاقواس. أو تحتوي على بيانات غير صحيحة أو غير كاملة أو قديمة أو قد تكون البيانات مكررة. قد يتضمن تنظيف البيانات انشطة اخرى مثل تنسيق البيانات وتوحيدها. الهدف من عملية تنظيف البيانات هو الحفاظ على بيانات ذات مغزى بإزالة العناصر التي قد تعوق تحليلها مما يؤثر على جودة النتائج حيث ان استخدام بيانات غير صحيحة أو غير متناسقة قد يؤدي حتما إلى نتائج خاطئة.

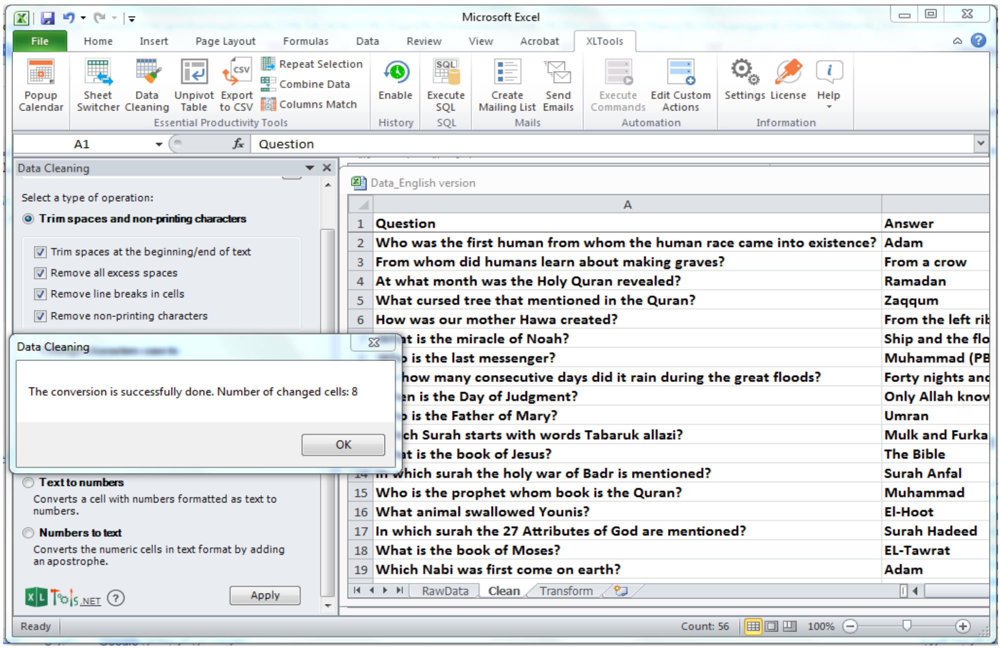
عند استيراد البيانات إلى اكسل أو لصق المعلومات إلى أوراق عمل من الإنترنت، قد ينتهي الأمر الى بيانات غير نظيفة، على سبيل المثال، قد تظهرفراغات زائدة بين الكلمات، و/ أو الرموز الغير قابلة للطباعة، الخ مما يتسبب في حدوث أخطاء عند المراحل المقبلة مثل تصفية البيانات، توحيدها، وتصديرها، الخ. يمكن تطبيق التصحيح اليدوي ولكنه قد يستغرق ساعات مضنية لتنظيف البيانات. خاصة اذا كان التعامل مع أوراق عمل كبيرة. بالإضافة الى ذلك قد لا يضمن التصحيح اليدوي اكتشاف و إزالة كل الأخطاء. ميزة تنظيف البيانات عن طريق أدواتاكسل يساعد على إزالة المسافات الزائدة وفواصل الأسطر قبل، بين وبعد الكلمات، حذف الأحرف غير القابلة للطباعة، تحديد كافة الخلايا الفارغة في وقت واحد ومن ثم معالجتها، تغيير النص إلى حالة سليمة وتحويل صيغ خلية من الأرقام إلى نص والعكس وذلك على نطاق واسع من الخلايا و بنقرة واحدة فقط.

كخطوة ثانية قمنا بنسخ البيانات من ورقة البيانات الاصلية لورقة عمل اخرى من مصنف اكسل ثم اجرينا عملية التنظيف عن طريق إزالة البيانات غير المتناسقة و البيانات التي لا نريدها في مجموعة البيانات ، حتى نتحصل على مجموعة نظيفة من البيانات:

لقد قمنا بإزالة المسافات الإضافية، و فواصل الأسطر والأحرف غير القابلة للطباعة وذلك بتحديد نطاق الخلايا التي تحتاج إلى تنظيف. ثم النقر فوق الزر "تنظيف البيانات" الموجود في علامة التبويب XLTools. ثم قمنا بتحديد العمليات المطلوبة: تقليم المسافات في بداية أو نهاية النص، إزالة كل المسافات الزائدة، إزالة فواصل الأسطر في الخلايا، و إزالة الأحرف غير القابلة للطباعة، ثم النقر على "تطبيق"، نلاحظ انه قد تم بنجاح تنظيف بعض الخلايا كما هو مبين في شكل رقم (1). بما ان الخلايا الفارغة قد تخلق الفوضى إذا لم تعالج مسبقا. فقد قمنا بتحديد كافة الخلايا الفارغة في وقت واحد ثم معالجتها. لتحويل تنسيق الخلايا من رقمية إلى نص، أو من النص إلى رقمية، يمكن تحديد العمليات المطلوبة: نص إلى أرقام، أو أرقام إلى نص. لإزالة البيانات المكررة قمنا بتحديد البيانات والذهاب إلى علامة التبويب بيانات، ثم إزالة التكرارات أو الصفوف المكررة. لتغيير حالة الاحرف في الإنجليزية وتوحيد تمثيل النص، يجب تحديد العمليات المطلوبة، الحالة السليمة، حالة الجملة، حروف صغيرة ، أو حروف كبيرة.

لمعرفة الأخطاء هنالك طريقتان لتسليط الضوء على الأخطاء في البيانات اما باستخدام التنسيق الشرطي أو عن طريق الانتقال الى الخاص. لقد طبقنا التدقيق الإملائي على مجموعة البيانات لتصحيح الأخطاء الاملائية ، حيث انه لا شيء يقلل من مصداقية العمل اكثر من الخطأ الإملائي. بما اننا استخدمنا عدة مصادر للحصول على البيانات، وحيث ان كل مصدر له تنسيقه الخاص به، فقد قمنا بحذف كل التنسيقات مرة واحدة ومن ثم تطبيق تنسيق موحد لكل البيانات. وحيث ان البحث والاستبدال لتنظيف البيانات في اكسل أمر لا غنى عنه فقد قمنا باستخدامه للبحث عن بعض الكلمات غير المناسبة ومن ثم استبدالها بأخرى اكثر تناسبا. عند الحصول على بيانات من قاعدة بيانات أو استيرادها من ملف نصي، قد يكتظ كل النص في خلية واحدة، في هذه

الحالة يمكن تحويل هذا النص إلى خلايا متعددة باستخدام وظيفة النص إلى العمود(text to Column)في اكسل.



Select a type of operation:

● **Trim spaces and non-printing characters**

☑ Trim spaces at the beginning/end of text

☑ Remove all excess spaces

☑ Remove line breaks in cells

☑ Remove non-printing characters

○ **Text to numbers**
Converts a cell with numbers formatted as text to numbers.

○ **Numbers to text**
Converts the numeric cells in text format by adding an apostrophe.

Apply

Data Cleaning

The conversion is successfully done. Number of changed cells: 8

OK

	A	
1	Question	Answer
2	Who was the first human from whom the human race came into existence?	Adam
3	From whom did humans learn about making graves?	From a crow
4	At what month was the Holy Quran revealed?	Ramadan
5	What cursed tree that mentioned in the Quran?	Zaqqum
6	How was our mother Hawa created?	From the left rib
	at is the miracle of Noah?	Ship and the flo
	is the last messenger?	Muhammad (PB
	how many consecutive days did it rain during the great floods?	Forty nights and
	en is the Day of Judgment?	Only Allah know
	is the Father of Mary?	Umran
	ch Surah starts with words Tabaruk allazi?	Mulk and Furka
	at is the book of Jesus?	The Bible
	which surah the holy war of Badr is mentioned?	Surah Anfal
15	Who is the prophet whom book is the Quran?	Muhammad
16	What animal swallowed Younis?	El-Hoot
17	In which surah the 27 Attributes of God are mentioned?	Surah Hadeed
18	What is the book of Moses?	EL-Tawrat
19	Which Nabi was first come on earth?	Adam

شكل رقم (1): استخدام xltools لتنظيف البيانات

3.3 تحويل البيانات

تحويل البيانات هو توحيد وتحويل البيانات إلى تنسيقات مناسبة للتعدين والمتوقعة من قبل أداة النمذجة. وتكون عادة من تنسيق نظام المصدر إلى التنسيق المطلوب لنظام وجهة جديدة. وتنطوي العملية المعتادة على تحويل الوثائق، ولكن تحويلات البيانات قد تنطوي أحيانا على تحويل برنامج من لغة كمبيوتر إلى اخرى لتمكين تشغيل البرنامج على منصة مختلفة. والغرض من هذه العملية هو اعتماد نظام جديد يختلف تماما عن النظام السابق. اكسل هو أداة عظيمة لاستخدامها عند الحاجة إلى أخذ البيانات في تنسيق معين ، ثم معالجتها لتحويلها الى تنسيق آخر، ثم دفع النتائج إلى اداة اخرى لاجراء معالجات اخرى. على سبيل المثال اذا اردنا تصدير ملف اكسل الى

تطبيقات اخرى تدعم تنسيق CSV،يمكننا تحويل ورقة العمل أولا الى تنسيق CSV ثم بعد ذلك تصدير ملف csv. الى تلك البرامج. يعمل اكسل بمثابة أداة تحويل للبيانات لنقلها من نظام إلى آخر، حيث انه يدعم العديد من التنسيقات.

بعد اجراء خطوات تنظيف البيانات تأتي مرحلة تحويل البيانات من تنسيق اكسل(xls.) الى التنسيق المطلوب. في هذه المرحلة قمنا أولا بنسخ البيانات التي تم تنظيفها مسبقا الى ملف اكسل آخر- حيث ان تنسيق CSV لا يدعم مصنف يحتوي على عدة أوراق عمل- ، بعد ذلك قمنا بتحويل ملف اكسل الى تنسيق ملفقيمة مفصولة بفواصل(CSV) وذلك باستخدام الامر حفظ باسم ثم اختيار تنسيق قيمة مفصولة بفواصل. ملفات CSVهي التنسيقات الشائعة لتبادل البيانات بين التطبيقات المختلفة. فهي مستخدمة على نطاق واسع لتخزين جداول البيانات (الأرقام والنص) كنص عادي فيمكن قراءتها باستخدام محررات النص القياسية [9]. ملفات تنسيق CSV تمكن من تحويل كميات كبيرة من بيانات الجداول بين العديد من التطبيقات و البرامج التي تدعم ملفات CSV، الامر الذى زاد من شعبيتها وقدرتها على البقاء، على الأقل كتنسيق بديل يمكن استيراده و تصديره، وعلاوة على ذلك فإن تنسيقات CSV تسمح للمستخدمين بتشخيص مشاكل البيانات فورا وبنظرة سريعة. بما ان ملفات CSV هي نص عادي، هذا يجعل فهمها سهلا سواء للمستخدم العادي أو حتى المبتدئ.

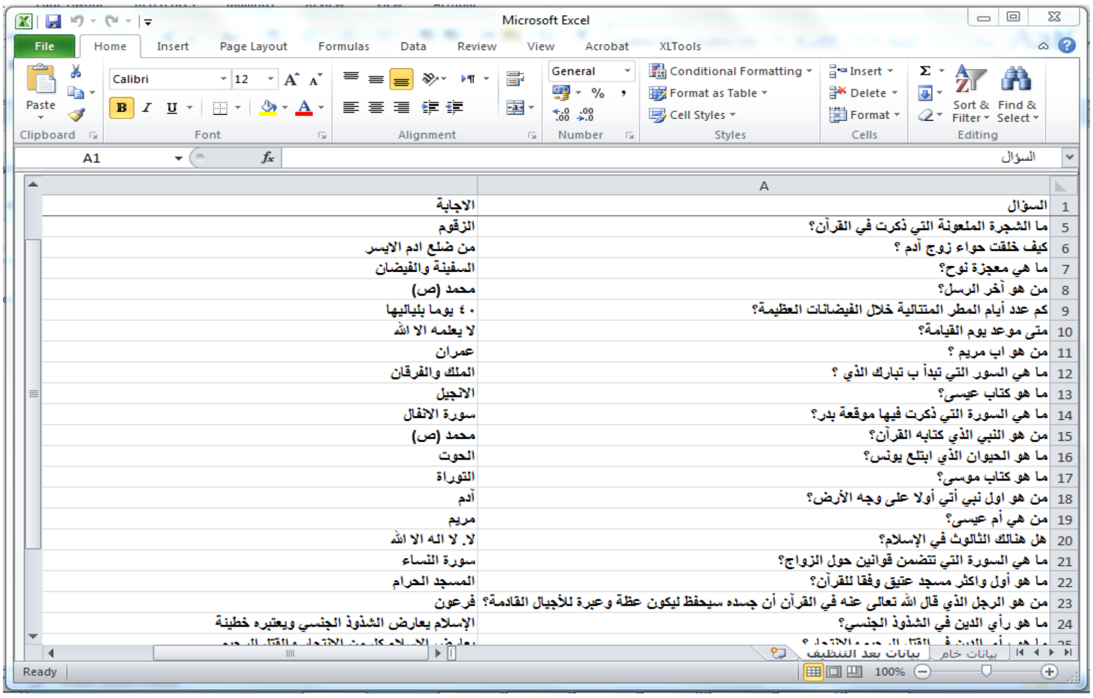
4 النتائج

تتكون مجموعة بيانات المجاميع(corpus) في مجملها ما يقرب من 1000 سؤال مع إجاباتها. تحتوي المجاميع (corpus) على بيانات نظيفة صالحة ذات تنسيق موحد يمكن استيرادها لأي تطبيق يدعم تنسيق CSV ومن ثم تحليلها. البيانات التي تم جمعها تحتوي على أنواع مختلفة من

الأسئلة كما هو مبين في جدول رقم (1). يبين الشكل (2) امثلة من أسئلة المجاميع (corpus) باللغة العربية بعد إدخالها بملف اكسل، كما يبين الشكل (3) امثلة باللغة الإنجليزية. نلاحظ اختلاف الاجابات من حيث الطول وذلك حسب طبيعة السؤال، حيث ان بعض الأسئلة قد تحتاج لمزيد من التوضيح. ونحن نتوقع أن هذه المجاميع (corpus) سوف تكون مفيدا في الدراسات التي تستخدم تنقيب أسئلة وأجوبة القرآن مثل التصنيف والتجميع الخ.

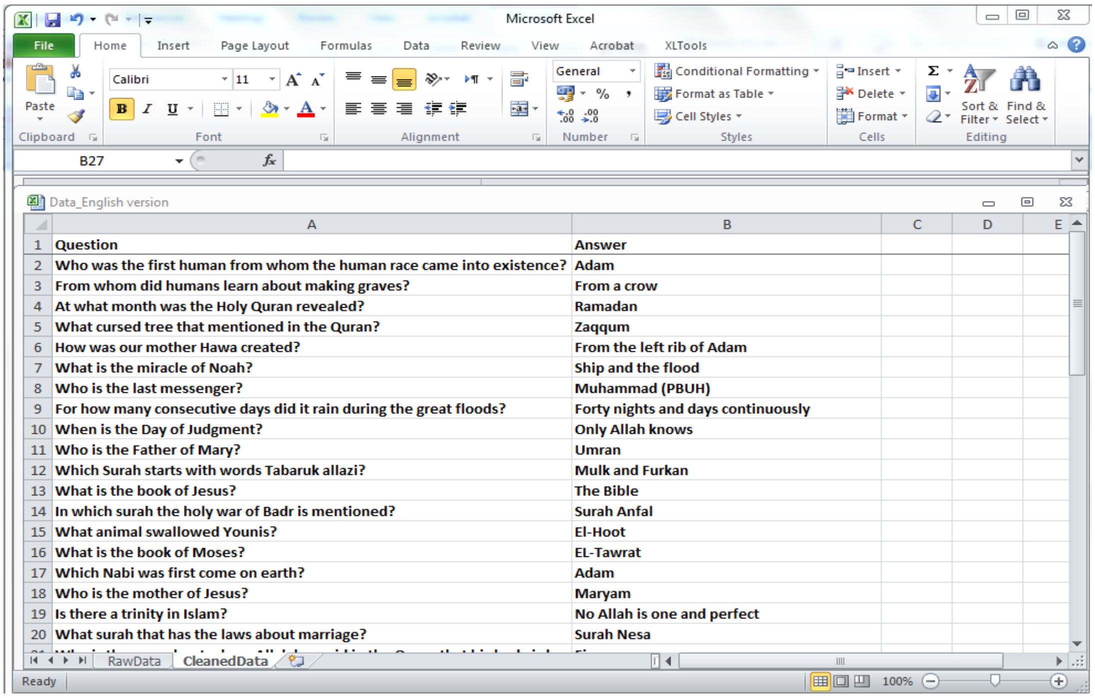
جدول رقم (1):أنواع الأسئلة

معرف النطاق	المعنى
ما (what)	يدل على سؤال يسأل عن الأشياء
من (who)	يدل على سؤال يسأل عن أشخاص
متى (when)	يدل على سؤال يسأل عن الوقت
اين (where)	.يدل على سؤال يسأل عن مكان
كم عدد (how many)	يدل على سؤال يسأل عن الأشياء القابلة للعد
لماذا (why)	يدل على سؤال يسأل حول السبب
كيف (how)	يدل على سؤال يسأل عن حالة أو موقف
اخرى (other)	جميع الأسئلة الأخرى



	A	
1	السؤال	الاجابة
5	ما الشجرة الملعونة التي ذكرت في القرآن؟	الزقوم
6	كيف خلقت حواء زوج آدم ؟	من ضلع ادم الايسر
7	ما هي معجزة نوح؟	السفينة والفيضان
8	من هو آخر الرسل؟	محمد (ص)
9	كم عدد أيام المطر المتتالية خلال الفيضانات العظيمة؟	٤٠ يوما بلياليها
10	متى موعد يوم القيامة؟	لا يعلمه الا الله
11	من هو اب مريم ؟	عمران
12	ما هي السور التي تبدأ ب تبارك الذي ؟	الملك والفرقان
13	ما هو كتاب عيسى؟	الانجيل
14	ما هي السورة التي ذكرت فيها موقعة بدر؟	سورة الانفال
15	من هو النبي الذي كتابه القرآن؟	محمد (ص)
16	ما هو الحيوان الذي ابتلع يونس؟	الحوت
17	ما هو كتاب موسى؟	التوراة
18	من هو اول نبي أتي أولا على وجه الأرض؟	آدم
19	من هي أم عيسى؟	مريم
20	هل هنالك الثالوث في الإسلام؟	لا. لا اله الا الله
21	ما هي السورة التي تتضمن قوانين حول الزواج؟	سورة النساء
22	ما هو أول واكثر مسجد عتيق وفقا للقرآن؟	المسجد الحرام
23	من هو الرجل الذي قال الله تعالى عنه في القرآن أن جسده سيحفظ ليكون عظة وعبرة للأجيال القادمة؟	فرعون
24	ما هو رأي الدين في الشذوذ الجنسي؟	الإسلام يعارض الشذوذ الجنسي ويعتبره خطيئة

شكل رقم (2): امثلة من البيانات باللغة العربية بعد تنظيفها



	A	B	C	D	E
1	Question	Answer			
2	Who was the first human from whom the human race came into existence?	Adam			
3	From whom did humans learn about making graves?	From a crow			
4	At what month was the Holy Quran revealed?	Ramadan			
5	What cursed tree that mentioned in the Quran?	Zaqqum			
6	How was our mother Hawa created?	From the left rib of Adam			
7	What is the miracle of Noah?	Ship and the flood			
8	Who is the last messenger?	Muhammad (PBUH)			
9	For how many consecutive days did it rain during the great floods?	Forty nights and days continuously			
10	When is the Day of Judgment?	Only Allah knows			
11	Who is the Father of Mary?	Umran			
12	Which Surah starts with words Tabaruk allazi?	Mulk and Furkan			
13	What is the book of Jesus?	The Bible			
14	In which surah the holy war of Badr is mentioned?	Surah Anfal			
15	What animal swallowed Younis?	El-Hoot			
16	What is the book of Moses?	EL-Tawrat			
17	Which Nabi was first come on earth?	Adam			
18	Who is the mother of Jesus?	Maryam			
19	Is there a trinity in Islam?	No Allah is one and perfect			
20	What surah that has the laws about marriage?	Surah Nesa			

شكل رقم (3): امثلة من البيانات باللغة الإنجليزية بعد تنظيفها

5 الإعداد للتحقيق ومواصلة التحليل

تحليل النص هو مجموعة من التقنيات اللغوية والإحصائية وتقنيات تعليم الآلة التي تسهم في استخراج المحتوى المعلوماتي الموجود في المصادر النصية. تحليل النص يشير إلى عملية اشتقاق معلومات عالية الجودة من النص، ويتم ذلك من خلال استنباط الانماط. تحليل النص ينطوي على استرجاع المعلومات، تحليل معجمي لدراسة التوزيع التكراري للكلمات، التعرف على الأنماط، العلامات / الشرح، واستخراج المعلومات، وتقنيات استخراج البيانات بما في ذلك تحليل الارتباطات، والتصور، والتحليلات التنبؤية. والهدف الأساسي من هذه العمليات هو تحويل النص إلى بيانات يمكن تحليلها، عبر تطبيق معالجة اللغات الطبيعية (NLP) والأساليب التحليلية. يستخدم هذا المحتوى المعلوماتي في العديد من المجالات مثل الأبحاث والدراسات أو التحليل الاستكشافي للبيانات.

يمكن أن يتم هذا التحليل باستخدام أدوات تنقيب البيانات أو أدوات تحليل النص. المهام النموذجية للتنقيب في النصوص تشمل التصنيف والتجميع، واستخراج المفاهيم، وتلخيص الوثائق، ونمذجة العلاقات بين الكيانات. من امثلة أدوات تحليل النص، آرغو(ARGO)، نووج (NOOJ)، ويكا (WEKA) الخ. يمكننا اجراء بعض الخطوات لتجهيز البيانات (preprocessing) استعداد لمزيد من التحليل، على سبيل المثال يمكن استخدام فلاتر ويكا(WEKA filters) لتجهيز البيانات وذلك لإزالة سمة (attribute) معينة أو إزالة (instances) و التي تلبي شرط معين، أو لتحويل السمات من النوع الذي فشلت WEKA في تحليله أوتنقيبه إلى نوع آخر صحيح يمكنها العمل عليه وذلك باستخدام الفلاتر NominalToString وStringToWordVector. بعد عملية تجهيز البيانات يمكن تطبيق تقنيات استخراج البيانات المختلفة.

6 قضايا عند جمع وإعداد البيانات

لقد واجهتنا عدة قضايا عند جمع البيانات وإعدادها، احدى هذه القضايا هومصدر اسناد البيانات: بعضها ليس لها إسناد التأليف. بعض المواقع نسخة من مواقع أخرى بحيث لا يمكن تحديد المصدر الحقيقي للنص. بالإضافة الى انه من الصعب التأكد ما إذا كانت هذه البيانات هي الأصلية أو مترجمة من لغات اخرى. هناك أيضا قضايا تتعلق بجودة الكتابة : بعض البيانات تحتوي على أخطاء إملائية، أوأخطاء نحوية مما يتطلب مزيدا من التدقيق اللغوي بالإضافة الى مشاكل التشكيل واحرف العلة القصيرة أو ظهور بعض الحروف الزائدة أو الرموز مما يسبب مشاكل عند تحليلها فيما بعد.

7 الخاتمة

قمنا بتجميع النسخة الأولى من مجموعة بيانات مجاميع(corpus)أسئلة واجوبة القرآن، استخدمنا في ذلك الطرق اليدوية و الآلية معا وذلك حسب مراحل جمع المجاميع (corpus): استخدمنا الطرق الآلية حيث ان الطرق اليدوية قد تستغرق ساعات مضنية أو قد لا تضمن اكتشاف و إزالة كل الأخطاء، و في الحالات التي تكون فيها الطرق الآلية غير مجدية، استخدمنا الطرق اليدوية حيث ان البيانات التي يتم جمعها يدويا تكون عادة ذات جودة عالية. لقد غطت المجاميع (corpus) مجموعة واسعة من أنواع الأسئلة. جمع البيانات يدويا يشكل تحديا كبيرا حيث ان الطرق الآلية لم تكن ناجحة دائما في تصفية المواد غير المناسبة أو غير المرغوب فيها.

نحن نخطط لتطوير وصقل هذه المجاميع و اضافة المزيد من الأسئلة في سياق مشروع لبناء التطبيقات التي تنطوي على الشرح والتحليل التلقائي لهذه المجاميع. في الختام، نحن نعتقد

إن إنشاء مجاميع مجموعة بيانات متكاملة و موحدة لأسئلة وأجوبة القرآن هو من الأهمية بمكان كمصدر مشترك لكل المهام التي تحتاج الى هذه البيانات بهدف تحسين الوضع الراهن، و ليتثنى سهولة استخدامها في الاختبار والتقييم في تطبيقات أبحاث القرآن أو أنظمة الرد على أسئلة القرآن من الإنترنت.

Appendix B stop words

NLTK English Stop Words

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers', 'herself', 'it', 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', 'should', 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', 'couldn', 'didn', 'doesn', 'hadn', 'hasn', 'haven', 'isn', 'ma', 'mightn', 'mustn', 'needn', 'shan', 'shouldn', 'wasn', 'weren', 'won', 'wouldn', 'What', 'Who', 'Why', 'While', 'Where', 'How', 'Which', 'Whom', 'When']

Example of Generated Format Python Arabic Stop Words File

#file name ../data/stopword0.6.csv)

```
STOPWORDS[u'لاسيما']=u'لاسيما';

STOPWORDS[u'ولاسيما']=u'و-لاسيما';

STOPWORDS[u'فلاسيما']=u'ف-لاسيما';

STOPWORDS[u'متى']=u'متى';

STOPWORDS[u'ومتى']=u'و-متى';

STOPWORDS[u'فمتى']=u'ف-متى';

STOPWORDS[u'أنى']=u'أنى';

STOPWORDS[u'وأنى']=u'و-أنى';

STOPWORDS[u'فأنى']=u'ف-أنى';

STOPWORDS[u'أي']=u'أي';

STOPWORDS[u'وأي']=u'و-أي';

STOPWORDS[u'فأي']=u'ف-أي';

STOPWORDS[u'أيان']=u'أيان';

STOPWORDS[u'وأيان']=u'و-أيان';

STOPWORDS[u'فأيان']=u'ف-أيان';

STOPWORDS[u'أين']=u'أين';

STOPWORDS[u'وأين']=u'و-أين';

STOPWORDS[u'فأين']=u'ف-أين';

STOPWORDS[u'بكم']=u'بكم';

STOPWORDS[u'وبكم']=u'و-بكم';

STOPWORDS[u'فبكم']=u'ف-بكم';

STOPWORDS[u'بما']=u'بما';
```

```
STOPWORDS[u'وبما']=u'و-بما';

STOPWORDS[u'فبما']=u'ف-بما';

STOPWORDS[u'أبما']=u'أ-بما';

STOPWORDS[u'أوبما']=u'أ-و-بما';

STOPWORDS[u'أفبما']=u'أ-ف-بما';

STOPWORDS[u'بماذا']=u'بماذا';

STOPWORDS[u'وبماذا']=u'و-بماذا';

STOPWORDS[u'فبماذا']=u'ف-بماذا';

STOPWORDS[u'بمن']=u'بمن';

STOPWORDS[u'وبمن']=u'و-بمن';

STOPWORDS[u'فبمن']=u'ف-بمن';

STOPWORDS[u'كم']=u'كم';

STOPWORDS[u'وكم']=u'و-كم';

STOPWORDS[u'فكم']=u'ف-كم';

STOPWORDS[u'كيف']=u'كيف';

STOPWORDS[u'وكيف']=u'و-كيف';

STOPWORDS[u'فكيف']=u'ف-كيف';

STOPWORDS[u'ما']=u'ما';

STOPWORDS[u'وما']=u'و-ما';

STOPWORDS[u'فما']=u'ف-ما';

STOPWORDS[u'ماذا']=u'ماذا';

STOPWORDS[u'وماذا']=u'و-ماذا';

STOPWORDS[u'فماذا']=u'ف-ماذا';

STOPWORDS[u'أماذا']=u'أ-ماذا';
```

```
STOPWORDS[u'أوماذا']=u'أ-و-ماذا';

STOPWORDS[u'أفماذا']=u'أ-ف-ماذا';

STOPWORDS[u'متى']=u'متى';

STOPWORDS[u'ومتى']=u'و-متى';

STOPWORDS[u'فمتى']=u'ف-متى';

STOPWORDS[u'مما']=u'مما';

STOPWORDS[u'ومما']=u'و-مما';

STOPWORDS[u'فمما']=u'ف-مما';

STOPWORDS[u'أمما']=u'أ-مما';

STOPWORDS[u'أومما']=u'أ-و-مما';

STOPWORDS[u'أفمما']=u'أ-ف-مما';

STOPWORDS[u'ممن']=u'ممن';

STOPWORDS[u'وممن']=u'و-ممن';

STOPWORDS[u'فممن']=u'ف-ممن';

STOPWORDS[u'من']=u'من';

STOPWORDS[u'ومن']=u'و-من';

STOPWORDS[u'فمن']=u'ف-من';

STOPWORDS[u'أنى']=u'أنى';

STOPWORDS[u'وأنى']=u'و-أنى';

STOPWORDS[u'فأنى']=u'ف-أنى';

STOPWORDS[u'أي']=u'أي';

STOPWORDS[u'وأي']=u'و-أي';

STOPWORDS[u'فأي']=u'ف-أي';

STOPWORDS[u'أيان']=u'أيان';
```

```
STOPWORDS[u'وأيان']=u'و-أيان';

STOPWORDS[u'فأيان']=u'ف-أيان';

STOPWORDS[u'أين']=u'أين';

STOPWORDS[u'وأين']=u'و-أين';

STOPWORDS[u'فأين']=u'ف-أين';

STOPWORDS[u'أينما']=u'أينما';

STOPWORDS[u'وأينما']=u'و-أينما';

STOPWORDS[u'فأينما']=u'ف-أينما';

STOPWORDS[u'حيثما']=u'حيثما';

STOPWORDS[u'وحيثما']=u'و-حيثما';

STOPWORDS[u'فحيثما']=u'ف-حيثما';

STOPWORDS[u'كيفما']=u'كيفما';

STOPWORDS[u'وكيفما']=u'و-كيفما';

STOPWORDS[u'فكيفما']=u'ف-كيفما';
```

PUBLICATIONS

Most of the research work done during the Ph.D has been presented in the following papers:

- Quran question and answer corpus for data mining with WEKA. In 2016 Conference of Basic Sciences and Engineering Studies (SGCAC) (pp. 211-216). IEEE.
- Using an Islamic Question and Answer Knowledge Base to answer questions about the holy Quran, International Journal on Islamic Applications in Computer Science And Technology, Vol. 4, Issue 4, December 2016, 20 -29
- تجميع مدونة اسئلة و اجوبة للقرآن الكريم (Compiling a Quran Question and Answer Corpus)، الدورة العاشرة للمؤتمر الدولي لعلوم وهندسة الحاسوب (ايكا ICCA) بالتزامن مع الدورة الثالثة للمؤتمر الدولي لتقنيات المعلومات والاتصالات في التعليم والتدريب:(تِسات TICET) مارس 2016م (ICCA'2016 International Conference on Computing in Arabic, 2016
- Evaluation corpus for restricted-domain question-answering systems for the holy Quran, International Journal of Science and Research (IJSR), Volume 6 Issue 8, August 2017