Question Processing for Arabic Question Answering System

معالجة السؤال في نظام سؤال جواب للغة العربية

By

Hani Maluf Al Chalabi

(Student ID number <u>120149</u>)

Dissertation submitted in partial fulfillment of the requirements for the degree of MSc in Informatics

Faculty of Engineering & IT

Dissertation Supervisor

Professor Dr. Khaled Shaalan

Dissertation Co-Supervisor

Dr. Santosh Ray

May-2015

DISSERTATION RELEASE FORM

Student Name	Student ID	Programme	Date
Hani Maluf Al Chalabi	120149	MSc. in Informatics	April, 2015

Title

Question Processing for Arabic Question Answering System

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that the content of this dissertation is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

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

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

I understand that The British University in Dubai may make that copy available in digital format if appropriate.

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

Signature

H.M.Al Chalabi

Abstract

Due to very fast growth of information in the last few decades, getting precise information in real time is becoming increasingly difficult. Search engines such as Google and Yahoo are helping in finding the information but the information provided by them are in the form of documents which consumes a lot of time of the user. Question Answering Systems have emerged as a good alternative to search engines where they produce the desired information in a very precise way in the real time. This saves a lot of time for the user. Question Answering systems are offered with the questions of natural language and proposed output is either the suitable answer recognized in a text or small text crumbs including the answer.

There has been a lot of research in the field of English and some European language Question Answering Systems. However, Arabic Question Answering Systems could not match the pace due to some inherent difficulties with the language itself as well as due to lack of tools available to assist the researchers. Therefore, in this dissertation, we will take the challenge to design and develop some modules of Arabic Question Answering Systems.

The task of Question Answering can be divided into three phases; Question Analysis, Document Analysis, and Answer Analysis. The part that our dissertation concern is the first phase, i.e., the Question Analysis phase. The question analysis phase consists of two major tasks namely Question Classification and Query Expansion beside other minor tasks such as stop word removal, Part of Speech tagging etc. We have proposed methods to accomplish these two major tasks in Question Analysis phase. We have used *Nooj* and Arabic WordNet (AWN) to implement our methods. In order to evaluate the performances of the proposed methods, we have used the corpus in Arabic language developed by Y.Benajiba which is available at http://users.dsic.upv.es/~ybenajiba/.

ملخص

نظراً لنمو سريع جدا للمعلومات في العقود القليلة الماضية، والحصول على معلومات دقيقة في الوقت الحقيقي أصبح من الصعب على نحو متزايد. محركات البحث مثل جوجل وياهو تساعد في العثور على المعلومات ولكن المعلومات التي تقدمها لهم هي في شكل وثائق الذي يستهلك الكثير من الوقت للمستخدم. انظمة السؤال والجواب برزت كبديل جيد لمحركات البحث حيث أنها تنتج المعلومات المطلوبة بطريقة دقيقة جدا في الوقت الحقيقي، وهذا يوفر الكثير من الوقت للمستخدم. وتقدم أنظمة السؤال والجواب مع أسئلة اللغة الطبيعية والإخراج مقترحاً هو إما إجابة مناسبة ضمن نص أو اجابة صغيرة.

هناك الكثير من البحوث في مجال اللغة الإنجليزية وبعض أنظمة السؤال والجواب الأوروبية، مع وجود قلة منها باللغة العربية، وذلك لوجود صعوبات لأنظمة السؤال والجواب باللغة العربية التي لا تتطابق مع وتيرة اللغة نفسها وكذلك بسبب عدم وجود الأدوات المتاحة لمساعدة الباحثين. ولذلك، وفي هذه الأطروحة، سوف نتخذ التحدي لتصميم وتطوير بعض وحدات اللغة العربية لأنظمة السؤال والجواب.

يمكن تقسيم نظام سؤال وجواب إلى ثلاث مراحل: تحليل المسألة، تحليل الوثيقة، وجواب تحليل. المرحلة الاولى هي الجزء الذي يتناوله موضوع الأطروحة، أي مرحلة تحليل السؤال. وتتكون مرحلة التحليل من مهمتين رئيسيتين هما تصنيف السؤال والاستعلام بجانب مهام بسيطة أخرى مثل إزالة كلمة توقف، جزء من خطاب علامات، الخ.

في مرحلة تحليل السؤال اقترحنا طرق لإنجاز المهام الرئيسية، حيث تم استخدام برنامج ادوات يدعى نوج وكذلك برنامج معاني الكلمات الشبكية باللغة العربية.

ولتقييم الانجاز التام من الطرق المقدمة فقد استخدمنا الإحضار في اللغة العربية المتقدمة والمطور من قبل ياسين بيناجيبا والمتوفر على الموقع **http://users.dsic.upv.es/~ybenajiba**

Dedication

Dedicated to my wife, children, and my parents

Acknowledgments

I would like to say special thanks to my leader and supervisor Professor Dr. Khaled Shaalan and Dr. Santosh Ray for their guidance, support, and motivation during my study. Their continuous support and advice helped me stay motivated and focused.

Thank you to all teachers and administration departments in BUiD for their support during my study.

Special thanks for my wife "Intesar" for her patience, support, and confidence in me during my study.

Special thanks for my great mother and father for their support and motivation in my life.

Table of Contents

Chapter One

Introduction

1.1 About Question Answering

In any Information Retrieval (IR) system, the method to get necessary information from the Internet is to enter some keywords in the system. By using specific query the IR system returns a list of possible related documents, where the user should scan to find the most suitable documents related to the information looking for. This method in some cases couldn't help the user to extract the relevant information efficiently from very big group of electronic documents, even though the structure of retrieval facility is easy. Question Answering (QA) is a technology that targets at finding the answer in large documents to the questions posed in natural language. Question Answering Systems (QASs) are fed with the questions in natural language and output is either the suitable answer recognized in a text or small text crumbs including the answer.

The difference between QA and conventional information retrieval as shown in Table 1.1, is that, in a QA the user is able to ask a question immediately in natural language to the system without any necessity to have query syntax. The system enables to answer the question in a form of extracting the exact answer from the documents. On the other hand, in IR, the input query is defined in the query language of the search engine. The output includes a ranked list of the documents supposedly containing the possible answers; the user then is responsible for reading the documents and find out relevant answers.

Table 1.1: The differences between Conventional IR and QA

	Conventional IR	**QA**
Input	Keywords	Natural language question
Output	A list of documents	Phrases and Words having the answer

1.2 Architecture of Question Answering System

A prototypical QAS consists of three main components; Question Analysis, Document Retrieval & Analysis, and Answer Analysis, Figure 1.1 shows the general architecture of a QAS.
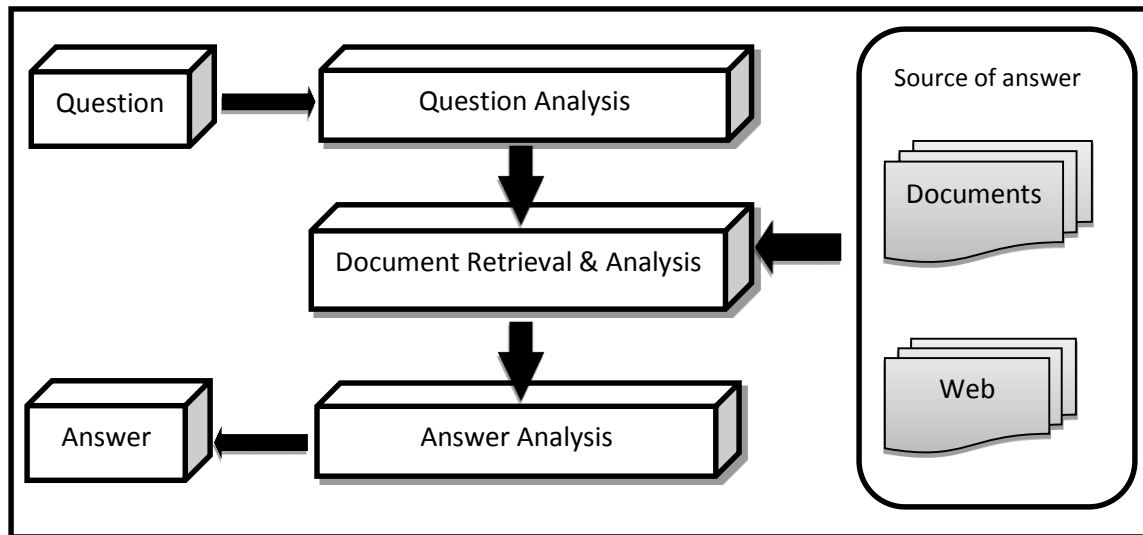

Figure 1.1: General architecture of a QAS

The method of the QAS design is usually known as the pipeline architecture of QAS. The questions flows from the first module "Question Analysis" to the end module which is the "Answer Analysis". Modules are sequenced such that the output of each module is an input to the module after (Tsur et al, 2004).

To analyze the question, "Question Analysis" module is responsible for analyzing the question and specifies what the question asking for, i.e. location, date, person's name etc. This module is capable to make an analysis as it contains a morphological analysis to determine the question class. A question class helps the system to classify the question type to provide a suitable answer (Zhang et al., 2003). This module may supplement additional keywords with the question.

The second module is the "Document Retrieval", which receives the classified question. This module depends on the identification of the subset components of the retrieval system which includes terms of assumed query from the collection of the total documents. The retrieval system returns the most likely documents that contain the

answer within a ranked list to be analyzed by the next sub-module which is "Document Analysis" (Hirschman et al, 2001).

The document analysis module takes the most likely answer list with the question classification description that shows what answer should be. This specification used to generate a number of answers which are closely related to the question to be sent to the "Answer Selection" module. This module selects the most correct answers among the phrases of certain type given by the Question Analysis (Ferret et al, 2001). The nominated answers which are chosen from the ranked documents in terms of the most correct answers are reverted to the user by this module (Dumais et al, 2002).

1.2.1 QUESTION ANALYSIS

The first step towards finding the answer is the analysis of the questions provided to the system in natural language. The main aim of Question Analysis is to understand the question purpose and meaning. To understand the question purpose, the question should be analyzed in different ways. Firstly, carry out the words' morpho-syntactic analysis in the question for English. This is done by tagging each word in the question as part of speech (PoS), in order to identify whether a word is singular noun, verb, plural noun, etc. After tagging the words, it is beneficial to find out the questioning information (what the question looking for). To understand the Arabic language question, it needs special handling to make systematic Natural Language Processing (NLP) systems. This is because the nature of Arabic words has been built from three or four roots of letters. The derivations of these words are shaped by adding the affixes (infix, prefix, and suffix) to each root depending on around 120 patterns (Abdelbaki et al, 2011),(Shaalan and Razza, 2009).

To get the question meaning, we need to classify the question type, which is the important step to get the actual answer. Question classification intends to group the question into pre-defined categories; Classification process is used to generate possible classes, which used to be predefined and limited with the word of key question. For example, a question can seek for date, time, location, and person. For instance, if system is able to understand the question "Who was the first American in space?" expecting that

the person name is in the answer, the search space of reasonable answers will be definitely reduced.

In general, almost all QASs involved a question classification module. The precision of Question Classification (QC) is very significant to the performance of the QAS. However, question classification is not a trivial task. Most of the systems help to comprehensive analysis of the question that determines more bounds on the answer entity. For instance; identify the question's keyword that helps in matching the sentences of candidate answer manner. Moreover, finding relations, syntactic, and semantic that must be hold between the entity of candidate answer and additional entities stated in the question (Rahman, 2015).

Many systems have been built a hierarchy for the question types according to the answer types, and enter the input question that suits a proper category in the hierarchy. Table 1.2 shows the question categories used to label the question type accordingly (Rahman, 2015). More details on Question Classification have been provided in chapter three.

Table 1.2: Question Hierarchy (Rahman, 2015)

Class	Class	Class	Class
ABBREV	Letter	Individual	**NUMERIC**
Exp	Other	Title	code
Abb	Plant	Description	count
ENTITY	Product	**LOCATION**	date
Animal	Religion	City	distance
Body	Sport	Country	money
Color	Substance	Mountain	order
Creative	Symbol	Other	other
Currency	Technique	State	period
Dismed	Term	**DESCRIPTION**	percent
Event	Vehicle	Definition	speed
Food	Word	Description	temp
instrument	**HUMAN**	Manner	size
Lang	Group	Reason	weight

There are different ways to categorize a question; the simplest way is to implement a pattern that matches the question to identify its type. Classification is more sensitive to

the sequence in which the pattern has been applied. For example, the patterns date-death and date-birth are implemented first, before the general pattern date (Suzuki et al, 2003). Once the question type being sought has been recognized, the question analysis remaining task has to recognize more constraints that the questions description type must meet as well. This process is simple as it needs only taking out the keywords from the remaining of the question to be used in finding the candidate answer sentences. These keywords may then be extended by using morphological and/or synonyms replacements (Srihari et al, 2001), (Shaalan et al, 2012)) or using query expansion techniques.

1.2.2 Document Retrieval and Analysis

Whatever the kind of QA architecture is selected; answering a question is almost includes some type of searching for retrieving documents that involves the answer (Navarro, 2014). QASs moved from classical document retrieval to Information Retrieval (IR) to save the user time by removing what needed to be searched via ranked list of documents to find the question answer. IR is away where question classification depends on within QA. In another way, IR is a task that retrieving documents relevant to a specific natural language query (Jurafsky et al, 2009). For instance, submitting a query by users via Google search engine and the query needs to be labeled or classified precisely to help in finding relevant documents (Rahman, 2015).

The function of IR is not to find real answers to the question but to recognize documents that involve the answer. The process of QAS within the step of IR is to take an upstream IR module related to excerpts relevant to documents from the corpus. This is done previously before proceeding to the extraction module of a downstream answer. In an attempt to locate relevant information accurately, the documents are dividing into passages, where they treated as documents. QASs also have a passage that go between the document retrieval and extraction components of the answer, which is called passage retrieval stage. The approach that based on passage retrieval is much easier to process QASs components than providing full document retrieval. It has an advantage of returning short text extracts as a replacement of full documents (Navarro et al, 2015).

Once the candidate documents or passages selected to get the answer, those may need further to be analyzed. At this stage, many of ways for document analysis needs to be considered, such as part-of-speech, splitting, tagging, and chunk parsing (recognizing some prepositional phrases, verb groups, noun groups, etc.) . To organize a clear link between a phrase of a particular type and the question, the following are often used: the pattern matching, syntactic structure, linear proximity, and lexical chaining (Gross et al, 2014). Ferret et al. (2001) identifies a QAS which depends on shallow syntactic analysis to recognize multiword terms with their alternatives in the documents. These documents have been selected to be re-ranked and re-indexed before the matching process against the representation of the question.

An extensive coverage statistical parser trained on the Penn Treebank is used to construct a reliance representation of the sentence in the answer documents (Harabagiu et al., 2001), (El Taher et al, 2014) . After that they match this reliance representation to be in the first logical order of the representation. Hovy et al. (2001) as well used the parser trained on the Penn Treebank but they considered generating a structure tree of syntactical oriented phrase. After that they match this into a representation of a logical form.

1.2.3 Answer Analysis

The final component in the general architecture of QAS is the representation of the user answer from the selected documents that includes the answer. The system that analyzes the question to get an expected answer follows some procedures to analyze the contents of the documents. These procedures can be done via the matching process which requires the text unit from the user answer text (in case splitting the sentence has been achieved) includes a string that its semantic type matches the expected answer (Toba et al, 2014). Same as previous components, there are some of ways to choose or rank the user answers. Moldovan et al. (2000) used an approach that, once the answer expression is found in the user answer paragraph, create a window of the answer around the user question. Different features like computing the whole score answer window through the word overlap between the answer window and the question used to be applied. For each user answer paragraph that includes the correct answer expression, a score has to be

6

derived for the answer window including the correct type. This score is considered for ranking overall user answers. Harabagiu et al. (2001) adds to this approach an extension; by applying machine learning algorithm to enhance the masses in the linear scoring function, which joins the features that characterizes the answer window.

Srihari et al. (2000) changed the order of the general approach by reversing it. This has been done by applying the question constraints more than the type of the expected answer as a filter to excerpt the suitable portion of the chosen sentences. On other side, they used for ranking the sentences such features like the number of unique keywords found in the sentence. The keywords order in the sentence used to be a comparison to their order in the question, and find out whether the keyword is verb or irregular matches.

Ittycheriah et al. (2001) has joint both predictable answer type matching with a set of word based comparison methods in one scoring function. They implemented this function on three sentences windows extracted from user answer documents. Light et al. (2000) delivered a discussion related to upper bounds on the comparison of word based approaches. Moreover, the frequency of user answer found to be measured as a standard for answer analysis and selection. This frequency represents the number of happenings linked to the question and it is also called redundancy answer selection (Clarke et al, 2002). This can be expanded to a larger set by counting the number of frequencies related to the set of documents that was delivered in the document analysis component (Dumais et al, 2002). Some QASs counts the number of answers occurs in the terms of the question from the whole document collection. Others go beyond the document collection by using the World Wide Web to catch the frequencies (Magnini et al, 2002).

1.3 Classification of Question Answering Systems

There are various dimensions along which QASs can be classified. Some of them are information retrieval techniques, query language, knowledge base for information, etc. (Voorhees, 2001). Accordingly, QASs can be classified into various groups as shown in Table 1.3.

Table 1.3: Question Answering Systems Classification

Dimension	Categories	Example Systems
Domain coverage	Restricted-Domain QASs	AQAS (Mohammed et al, 1993)
	Open-Domain QASs	ArabiQA (Benajiba et al, 2007) QASAL (Brini, 2009) AQUASYS (Bekhti et al, 2011)
Information retrieval technique	Rule Based QASs	AQUASYS (Bekhti et al, 2011) QASAL (Brini, 2009)
	Statistical QASs	ArabiQA (Benajiba et al, 2007)
Answer source	Automated QASs	QASAL (Brini, 2009) AQAS (Mohammed et al, 1993) ArabiQA (Benajiba et al, 2007) AQUASYS (Bekhti et al, 2011) QARAB (Hammo et al, 2002) DefArabicQA (Trigui, 2010)
	Collaborative QASs	N/A
	Hybrid QASs	N/A
Knowledge base	Web-based QASs	QARAB (Hammo et al, 2002) DefArabicQA (Trigui, 2010)
	Corpus-based QASs	AQAS (Mohammed et al, 1993) ArabiQA (Benajiba et al, 2007)
	Hybrid QASs	QARAB (Hammo et al, 2002) DefArabicQA (Trigui, 2010)
Language supported	Monolingual QASs	AQAS (Mohammed et al, 1993)
	Multilingual QASs	ArabiQA (Benajiba et al, 2007)

a) Domain Coverage: QASs are classified into two categories depending upon the domains covered by them:

- *Closed-domain* question answering treats with questions beyond a particular domain (for instance: automotive maintenance or medicine), and can be shown as a calmer task that because of Natural Language Processing (NLP) systems can feat domain-specific knowledge regular officially in ontologies. Consequently, *Closed-domain* might denote to a state that only a limited question types are recognized; like

questions requesting for expressive rather than technical information (Hovy et al, 2001).

- *Open-domain* question answering treats with questions beyond almost anything, which only rely on world knowledge and general ontologies. Moreover, the systems usually have more data exist from which to get the answer (Hovy et al., 2001).

b) Information Retrieval Technique: Different techniques used by different QASs to retrieve answers for a given question by the user. These techniques can be divided into two categories: Rule Based Systems and Statistical Systems (Li and Roth, 2004).

Rule Based QASs characteristically do the detailed QA by recognizing user passages. This can be done by using IR techniques and applying syntactic/semantic parsing for matching passages to find particular answers for the user question (Li and Roth, 2004). On other hand, *Statistical systems*, take the advantage of the statistical patterns by text in order to retrieve the answer. For example; keywords based retrieval used to call a part of documents that may possibly contain the answer. Moreover, Statistical systems use to capture the semantic functions by considering the texts like bags of words. The semantic representation allows QASs to find the answer for more complex queries (Mahendra et al, 2008).

c) Answer Source: In field of QASs, most of researchers focused on *automated* retrieval answer for a given question by the user. An automated QAS receives a question from the user in some natural language; this can be done by applying several techniques of information retrieval. In addition, produces the answers automatically in form of passages, sentences or documents link (Carpineto, et al, 2012). In *Collaborative* QA (CQA) systems; to retrieve an answer for a given question by the user, it needs a matching of the question to the best answers. A key to improve the CQA systems is to support askers with effective and cooperative service by reducing the time needed to provide the user's answers. It requires re-ranking a list of answers that includes the topic of the answer given to the user in effectively and efficiently (Liu et al, 2011). While in *Hybrid* QA systems, the answer used to be taken depends on the output of answer validation techniques and QA based systems. This can be done by combining the question and the answer option to generate a Hypothesis. Stop words used to be removed

from the Hypothesis. In addition; query words used to be identified to find most relevant sentences in matched documents. Combining both high scores of Hypothesis and matched sentences gives the most probable answer to the user's question (Pakray et al, 2001).

d) **Knowledge Base**: With regard to the knowledge base, QASs differ from which the answer used to be retrieved. QASs like Webclopedia (Hovy et al, 2001), use as knowledge base a local corpus while others use their knowledge base as the World Wide Web to retrieve relevant information. This system differs from Wikipedia which represents a group of documents established by experts. Wikipedia is a multilingual QAS which joined rule based approaches and statistical techniques to get answers (Giampiccolo et al, 2008).

e) **Language Supported**: QASs can be found either as monolingual QASs or multilingual QASs. Monolingual QASs backup the user questions using one language only and then return the related sentence or passage in same language. It processes the documents using the user question language without performing any translation. On the other hand, Multilingual QASs take more than one language from the user query and translate them into multiple language queries, then retrieve the documents that include these queries. Finally, give back the user the answers in a corresponding language (Magnini et al, 2004).

1.4 Research Questions

There are several issues in designing and development of Arabic Question Answering Systems such as efficient question classification, use of semantic resources in query expansion, efficient ranking of documents etc. This dissertation attempts to answer the following research questions:

The first research question is:

- Can the Precision of retrieving the answer be improved by efficient question classification?

– If yes, how to design efficient methods for question classification?

The second question of this research is;

- How the use of sematic web in query expansion process can help in increasing the Recall of the Question Answering System?

1.5 Research Objectives

The main objective of this research is to design and develop Question Analysis (QA) module of Arabic Question Answering Systems.

To achieve this objective, this research proposes to accomplish the following tasks:

1. Question classification: To design a question classification method to classify questions into predefined classes and identify the entities expected to be present in the answers.

2. Query Expansion: To expand the user queries using semantic resources to increase the Recall of the document retrieval phase.

1.6 Contribution of the Dissertation

QASs as mentioned in section 1.3 are classified into two categories depending upon the domains covered by them: Closed-domain and Open-domain.

To analyze the question using QAS, we need to generate a representation for the information, which can be done in the Question Analysis component of QAS (as mentioned in section 1.2.1). It consists of various sub-processes, such as question classification, derivation of expected answer, keyword extraction, and query expansion.

This dissertation mainly focuses on the question analysis component of a QAS. First, *Question Classification* represents the main part of QAS nevertheless of numerous types of architectures. Furthermore, it has been observed that the pursuance of question classification has made an important influence on the pursuance of QAS (Al Chalabi et al, 2015). In fact, there are two incentives for question classification; *predicting the answer entity*, and *developing answer pattern* (more details in chapter three).

Recently, there has been a lot of research in the field of English and some European language QASs. However, Arabic QASs could not match the pace due to some inherent difficulties with the language itself as well as due to lack of tools available to assist the researchers. Therefore, in this dissertation, we took the challenge to design and develop some modules of Arabic QASs. Moreover, chapter three of this dissertation shows various question patterns to fully characterize the Arabic questions in a format of machine processable using a natural language tool "*Nooj*" used to design and test the patterns (Al Chalabi et al., 2015).

Second, Query Expansion known as the addition of the related terms, has displayed its efficiency in enhancing the results of information retrieval process. Millions of users are frequently searching over Internet using keywords by connect with their information needs. Unfortunately, the queries most possibly get failed because of word mismatch between the user query and document lexicon. The best ways to solve this problem is to extend the query by using a thesaurus which shows the semantic connection (relationship) between terms of the query.

1.7 Outline of the Dissertation

The remaining chapters of this dissertation are arranged as follows:

- Chapter Two: This chapter presents a general approach to QA in detail and outlines earlier research on QASs with specific assertion on the task of TREC-QA, Web-based QA, and Arabic QA. It looks as well at the related QA techniques and how research for QASs took place in these areas.

- Chapter Three: The chapter concerns of showing the definition of Question Classification, its methods, proposed methods for Arabic Question Classification (including evaluation, tools, and corpus), and results.

- Chapter Four: The chapter concerns of showing the definition of Query Expansion, its methods, proposed methods for Arabic Query Expansion (including evaluation, tools, and corpus), and results.

- Chapter Five: The chapter concludes the work in this dissertation and adds remarks for developing methods related to components of Arabic QASs.

Chapter Two

Literature Review

2.1 The History of Question Answering

The growth of systems that react with the natural language users was the goal for the research community of the Artificial Intelligence. Since 1960s, when the field was at beginning, a set of database related to the natural language have been created like; dialog systems, language understanding systems, and front-ends. Simmons (1965) illustrates at the primary studied, not less than fifteen applied English language QAS built over the earlier five years than his paper. The idea of a QAS was born in 1950 when Turing (1950) proposed a chore known as "Imitation Game" which then became famously known as "Turing Test" where a human can communicate with a machine through an interface (teletype) which can as questions of it. Turing can be considering as a machine when a human could not make a difference between a responses of the machine and another human which already communicating through teletype.

In early 1960s, there was an interest in improvement for database query systems related to the natural language front-ends, like BASEBALL system, which was the better early QAS and most fruitful system (Green et al, 1961). BASEBALL is a program that built specializes for baseball games which include answering questions related to the statistical of the played baseball games in the American League for one season. The program was able to answer questions using techniques of shallow parsing over the natural language query, the techniques used in order to classify the statistics and teams in the questions. It was also capable to control some extensive queries that include collecting data available in different records related to the baseball database and then return the suitable answer (Green et al, 1961). Another system uses the same techniques found in the earliest known as ELIZA (Weizenbaum, 1966).

Moreover, the most well-recollect other previous work in this period is the LUNAR system, which capable to access to the information that includes two databases for the samples of the moon rock. LUNAR was planned "to allow a lunar geologist to suitably access, compare and evaluate the data of the chemical analysis on lunar rock and soil

arrangement that was collecting as a result of the Apollo moon mission" (Woods, 1973). The system was operated by interpreting the questions of natural language queries into the query language of the database engine.

However, these systems were excellent in responding to the certain classes of the questions, the systems were unable of replying to any questions (natural) that might propose themselves within a dialog with the user.

Early on, in the 1970s, the famous area for QA that engrossed researchers to submit their applications was natural language database and front-ends. QA in human-machine dialogue was another area of primarily theoretical attention, where such dialogue systems have been built to assist researchers to understand the concepts of modelling the human dialogue. For instance; SHRDLU (Winograd, 1972) and GUS (Bobrow et al., 1977), both represented a notable capacity to learn natural language.

SHRDLU was a system that has been built to simulate the robot capable of moving objects in the world of blocks. GUS has been built to simulate a travel consultant which had an access to the limited database of information related to the airline flights. In 1977, Lehnert developed a system called QUALIM which was the most notable work. The system showed the procedure of QA as unique as both understanding and answering the questions depends on the story context as well as the realistic concepts of suitability of answer (Lehnert, 1977). More work has been applied in story comprehension, many of it through psychology community (Kintsch, 1998) and the work of improving computational models of understanding the story via 1980s and 1990s. However, an interest has been registered in the area that follows the designation of a task of reading comprehension evaluation (Hirschman et al, 1999).

Since the end of 1990s, the interest of QA became remarkable increased among the research community especially in the natural language QA when the track of the introduction of QA began with TREC-8 in 1999 in the conferences of text retrieval (more details in section 2.2.2 of this chapter) (Hirschman et al, 1999).

2.1.1 Contributions in individual components of QASs

Recently, in each evaluation task of QAS, several systems of QA were submitted, such as TREC, CLEF and NTCIR etc. This happened especially in years 2002, 2003, and 2004

where the number of research groups participated in QA track as follows; in 2002 were 34 and 16 in 2003 as well as 28 in 2004 (Tsur et al, 2004). Each group applied their private system, where different architectures and techniques have been used such as; 1) Database of external knowledge, 2) Question type ontology, 3) Generation of answers, 4) Heuristics for extracting answers of certain types, 4) Answer justification, 5) Feedback loops, 6) Inference rules, 7) Even logical analysis, and 8) Machine learning. Thus it's difficult to collapse all these techniques in unique architecture. However, the approach of general QA is limited by the environment of the task itself which enable us to have a general architecture of a prototypical QAS. As an optimum method of QAS, the system should be able to understand what being asked and also find the proper and relative documents or passages belong to the question to find suitable answers. Meanwhile, systems require specifying the amount of answers and select the best to be viewed by the user. As we mentioned in chapter one, the QAS (prototypical system) consists of four main components; these are Question Analysis, Document Retrieval, Document Analysis, and Answer Selection.

The method of QAS usually known as a pipeline of QAS, where the questions flow from the first module of which is "Question Analysis" to the end module which is the answer. Modules are sequenced such that the output of each module is an input to the module after (Tsur et al, 2004).

To analyze the question, the first module which is called "Question Analysis" is responsible on analyzing the question and specifies what the question asking for, i.e. location, date, person's name etc. This module capable to make an analysis as it contains a morphological analysis to determine the question class. A question class helps the system to classify the question type to provide a suitable answer; this might needs more clarification from the user (Zhang et al., 2003).

The second module is the "Document Retrieval", which receives the classified question, this module depends on the identification of the subset components of the retrieval system which includes terms of assumed query from the collection of the total documents. The retrieval system returns the most likely documents that contain the answer within a ranked list to be analyzed by the next module which is "Document Analysis" (Hirschman et al, 2001).

The document analysis module takes the most likely answer list with the question classification description that shows what answer should be. This specification used to generate a number of answers which are closely related to the question to be sent to the "Answer Selection" module. This module selects the most correct answers among the phrases of certain type given by the Question Analysis (Ferret et al, 2001). The nominated answers which are chosen from the ranked documents in terms of the most correct answers are reverted to the user by this module (Dumais et al, 2002).

2.2 Question Answering at TREC Task

Text REtrieval Conference (TREC) is one of the annual conferences that arranged around the world to support the environment for the relative evaluation of different information retrieval systems from commercial and academic research groups (Moldovan et al, 2000). It has been established in 1992, and has developed significantly from its original form. In 1999, the 8th conference of TREC (TREC-8) proposed at first time the track of QA which needed to answer factoid questions by getting a text clip which includes an answer to the question. The track of TREC QA supplies a group of documents (numerous gigabytes of articles and newspapers from different sources) and a group of questions that needs short fact-based answers (500 questions for TREC-2003) (Harabagiu et al, 2003).

In 1999 it was guaranteed that the group of documents has an answer of each question, this done by TREC-8 and TREC-9 (Moldovan et al, 2000). In TREC-2001, the system here had a responsibility to recognize the answer rather than replying a misleading answer (Ellen, 2001). The QA track for TREC-2002 has two tasks; the list task and the main task, the task required by the system was to give the exact answers and all members were restricted to make one answer for each question not five like previous TRECs (Soubbotin et al, 2002).

For QA track of TREC-2003 includes two tasks; the main task (includes three kinds of questions; definition questions, list questions, and factoid questions) and the passages task (answer factoid questions by getting a text clip which includes an answer), it is important to give the evaluations of the first large questions (list and definition) (Harabagiu et al, 2003). In TREC-2004 a set of series' questions used to be defined as a set of targets, where both factoid and list question are not separate, as a replacement, all

are related to specified targets. More members are included in resolving the tasks of both list and definition question (Guo, 2004).

In TREC-2005, the task that used a classic ad hoc retrieval with the emphasis on the effectiveness of a unique topic rather than average one, this track known as "Robust Retrieval Track"[1]. In 2006 another track has been announced in the QAS TREC-2006 called "Terabyte Track"; the aim of this track is to discover whether/ how the Information Retrieval (IR) community able to scale the traditional IR to test a collection of based evaluation to the large document collections[2]. In TREC-2007 a Question Answering track has been used as a QAS, this track designed to get a step nearer to IR rather than document retrieval[3]. In TREC-2009 a track used was "Million Query Track", the aim of this track was to exam the hypothesis that generated from topics that incompletely judged rather than a tool used to build a collection of traditional TREC pooling[4]. In TREC-2010, a Blog Track has been announced in QAS, the goal of this track was to explore the behavior of information seeking in the blogosphere[5]. In TREC-2011, an Entity Track that has been run as the QAS, this track aim was to implement the tasks of the entity-oriented search in the World Wide Web. The tasks based on returning particular objects instead of any type of document[6]. In TREC-2012, a track that runs at that time was "Legal Track" which its goal was to develop the search technology that recognizes the lawyers' needs to be involved in the effectiveness of discovering the collections of digital documents[7]. In TREC-2013, a track called "Crowdsourcing Track" has been announced to discovering developing crowd-based approaches for searching evaluation as well as improving hybrid automation search systems[8]. Recently an example of one of the tracks used in TREC-2014 is "Session Track" which its aim is to supply the required resources in a way of testing the collections that can help to evaluate the IR system utility by simulating the

---

[1] http://ciir.cs.umass.edu/research/hard/

[2] http://www-nlpir.nist.gov/projects/terabyte/

[3] http://trec.nist.gov/data/qa/t2007_qadata.html

[4] http://ciir.cs.umass.edu/research/million/

[5] http://trec.nist.gov/tracks.html

[6] wiki.ir-facility.org/index.php/TREC_Chemistry_Track

[7] http://trec-legal.umiacs.umd.edu

[8] https://sites.google.com/site/treccrowd

user interaction. This can be done by a series of user interactions and queries instead of a query that includes a single "one-shot"[9].

2.3 Related Work on the Web-based QA Systems

Recently, most of the QASs have been researched widely and achieved important development in performance. However, the perfect source of answers can be taken via the Web due to the availability of great amount of information online. The researchers of QA tried to discover a range of the Web usages to focus existing techniques of QA in the context WWW. Many systems of QA are publicly available on the Web, like; QuASM[10], START[11], SiteQ/E[12], AskJeeves, IONAUT[13], LCC-Web[14], Encarta[15], AnswerBus[16], AskMAR[17], etc.

Among these systems, we highlight some of the representatives, STRAT (SynTactic Analysis using Reversible Transformations), this is the first Web-based QAS that has been an on-line effective since December, 1993. It has been developed by Boriz Katz using a key technique known as "Natural Language Annotation" which enables the systems to answer questions using a pre-compiled knowledge base (Katz, 1997).

AnswerBus is a QAS which is an open domain based on a level of sentence that retrieves Web information. It can receive natural language questions from the users in Multi-lingual and produces matched answers that can be accessed via the Web. Five types of search engines as well as dictionaries took place to retrieve Web pages that includes answers, like; Google, Yahoo, AltaVista, WiseNut, and Yahoo News. AnswerBus[18] produces sentences which include answers, it shows that practical QA is highly feasible on the Web (Zheng, 2002).

---

[9] http://ir.cis.udel.edu/sessions
[10] http://www.espressoenglish.net/an-easy-way-to-form-almost-any-question-in-english/
[11] http://start.csail.mit.edu/index.php
[12] http://ressell.postech.ac.kr/~pinesnow/siteqeng/
[13] http://ilk.uvt.nl/~antalb/tint/week3/io_index.html
[14] http://www.languagecomputer.com/
[15] http://microsoft-encarta1.software.informer.com/
[16] http://www.answerbus.com/index.shtml
[17] http://www.askmar.com/
[18] http://www.answerbus.com/index.shtml

LCC's QAS has been improved by Language Computer Corporation (LCC), in "TREC-QA 2002", "TREC-QA 2003", and "TREC-QA 2004" was the best QA system (Harabagiu et al., 2003). LCC has been achieved because of the technique used to combine the intensity of Information Extraction (IE) with the "expansion of axiomatic knowledge" illustrations generated from WordNet in order to justify answers that are extracted (Harabagiu et al., 2003).

Dumais et al, illustrated a QAS which is an open-domain Web called "AskMAR" (Brill et al., 2002) that stratifies simple query (rewriting) to the clips given by Google as well as a group of 15 handcrafted filters (semantic) to reach a good accuracy.

Furthermore, other web-based QASs were resulted in a different development, like; AQUAINT, Textract, IBMPQ, Aranea, DIOGENE, QUANTUM, QALC, QALC, Tequesta, etc. (Zhang, 2004).

2.4 Related Work on the Restricted Domain QA Systems

There is a long history for QASs starting from the systems over databases working, like; BASEBALL (Green et al, 1961) and LUNAR (Woods, 1973). As illustrated before, the QA trend is to constraint on open-domain, this is driven by the Track of TREC-QA. Nevertheless, the QAS with open-domain techniques has been lacking to give the specific domain for all question types. This is because no limitation has been given on the question type or on the specific vocabulary of the user, as well as it is hard to build an ontology common knowledge base (Harabagiu et al, 2003).

Many researchers have been built an advanced limited domain QASs. Tsur et al. (2004) produced a biographical QAS called "BioGrapher" which capable to address the problem using algorithms of machine learning for biography classification. Biographer used to collect answers by searching in a given biographies collection using methods custom made for the limited domain nature as well as finding the answer from the Web search engine incase the answers couldn't be found.

Benamara (2004) illustrated an experiment for the domain of tourism to produce a logical based QAS called "WEBCOOP", which integrates the representation of knowledge and the methods of advanced reasoning, this is done to produce a cooperative responses to the queries of natural language on the Web. Nguyen et al. (2004) stated a limited domain of

QAS with semantic information that includes finding a group of specific terms and constructing the hierarchy of the concept which can successfully characterize the significance of retrieving the candidate to its consistent question. In a large company, these services can be offered for replying questions. Niu et al. (2004) delivered a QA research for the texts of clinical-evidence, which able to classify incidences of the semantic classes ("patient outcome", "medication", and "disease") as well as the relationship between them to specify an outcome whether it is negative or positive.

Chung et al. (2004) defined a research which works for QA by limiting the domains of the question and getting answers from the documents that are semi-structured or pre-selected on the Internet.

2.5 A Review of Tools on Question Answering Systems

Question Answering System in IR and NLP is an automatic task that provides an answer for a given question by a user. Arabic language differs from other languages in its richness that needs special handling to make systematic NLP systems (Abdel-Monem et al , 2008). As we mentioned in chapter one, this is because Arabic words built from three or four roots of letter, and derivations of them are formed by adding to each root the affixes (infix, prefix, and suffix) depending on around 120 patterns (Abdelbaki et al, 2011),(Shaalan and Razza, 2009). In Arabic language, the derivations are almost like; lemma = root + pattern. Figure 3.1.
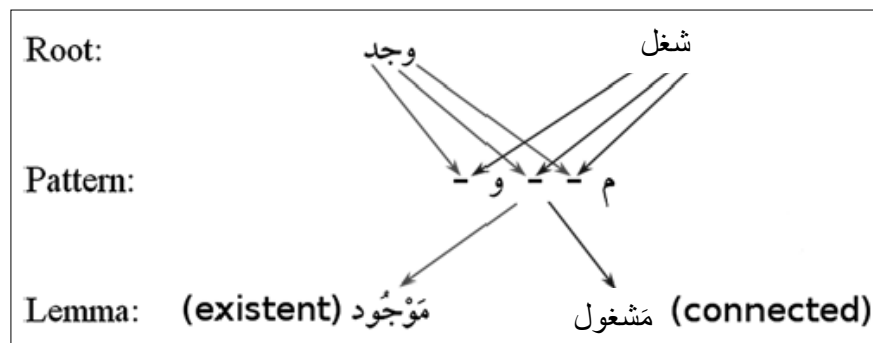

Figure 2.1: Arabic Derivation

The lack of diacritics is one of the Arabic difficulties posed in Modern Standard Arabic (MSA), which added to the confusion in question (Rashwan et al. 2011) , (Shaalan, 2014) . For instance; the word "جلد" has a meaning in MSA "جِلد" (skin) or "جَلَدَ" (hit somebody)

depending on the context. MSA can solve the problem of the Arabic discretization in a rate of small error about 3.1% to 12.5%.

Arabic QA has some tools provided by researchers to solve the problems of Arabic NLP, such as; AraMorph[19], MADA+TOKAN, AMIRA Tools, Fassieh, Morphological Analyzer, Nooj, Stanford NLP, etc.

AraMorph is a morphological analyzer which is the Buckwalter Arabic Morphological Analyzer (BAMA). It translates the Arabic words based on the transliteration system called Buckwalter, this system has a dictionary which based on Arabic stemmer. For instance the word كتاب can be transliteration to ktAb according to the morphological analysis. In general, adding unneeded execution to the stemming process cause an affect and making it slower (Buckwalter, 2002). Habash et al. (2009) designed MADA+TOKAN a toolkit which is available free and offers different services of Arabic NLP like; diacritization, tokenization, morphological disambiguation, stemming, Part of Speech (PoS), and lemmatization.

MADA examines the probable analysis for each word, then choose the suitable analysis by matching it to the current context. This can be done by using a model classification called Support Vector Machine (SVM). TOKAN works interchangeably by taking MADA output and generates the tokenized output in a suitable format.

Diab (2009) introduced a toolkit called AMIRA which involved "PoS tagger, clitic tokenizer, and base phrase chunker (shallow syntactic parser)". The AMIRA technology based on managed learning without any dependency on obvious modeling or the knowledge that has deep morphology. AMIRA can give the user a flexible to ask for tokenized PoS tagged or non-tokenized output.

Benajiba et al. (2007) used Named Entity Recognition (NER) depending on combining Maximum Entropy model with PoS tagging information, they used a system called "ANERsys" ver. 1.0. In same year (2007) Benajiba and Rosso designed the second version of ANERsys (ver. 2.0) in order to solve the problems of having multi-token (Benajiba and Rosso, 2007). Another important and effective maximum entropy based named entity recognizer for Arabic documents is NERA (Shaalan and Raza, 2009), (Shaalan and Oudah, 2014).

---

[19] http://www.nongnu.org/aramorph/

Nooj[20] written by the Professor Max Silberztein in 2002, it is an engineering development environment of a freeware linguistic. It is written by C#.Net programming language on the Visual Studio .NET Platform.

Inside Nooj we can use any lexical, morphological, syntactic, and semantic information in the text. Figure 3.2 (more details on Nooj will be discussed in chapter three of this dissertation).
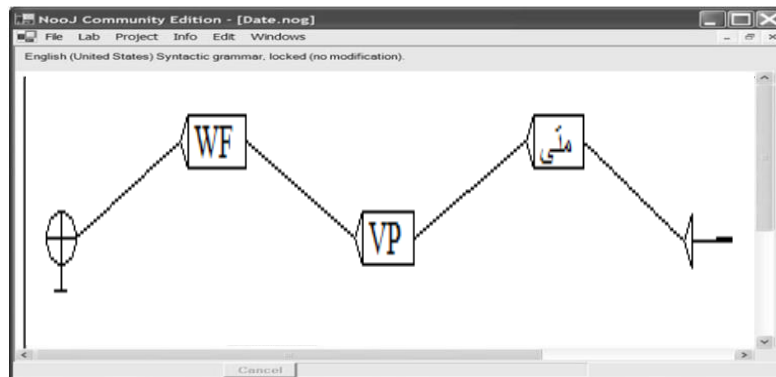

Figure 2.2: Nooj using NLP Tools

Stanford[21] NLP is a toolkit which consists of a group of libraries and can be used in QA tasks, as well as it covers the most shared tasks of NLP. Some libraries that Stanford consists of are; Stanford Parser, Stanford PoS Tagger, and Stanford Named Entity Recognizer. See Figure (2.3)

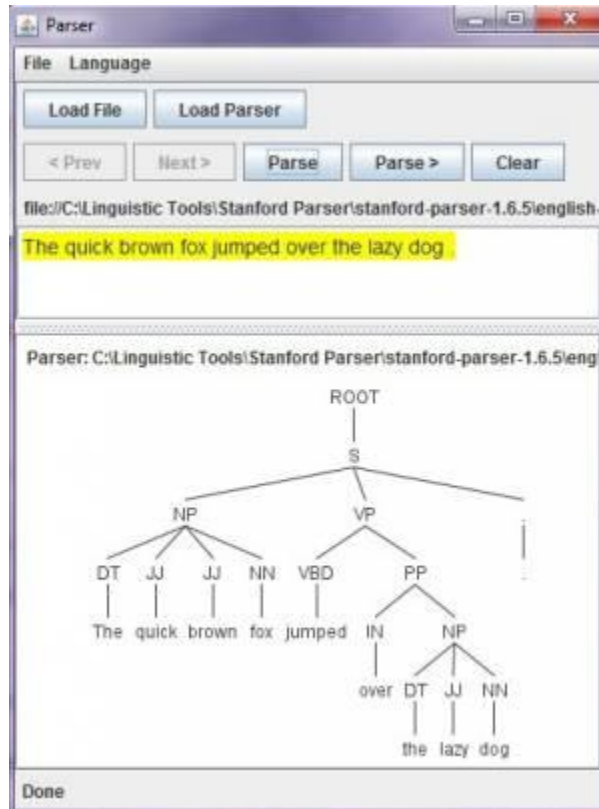[20] http://www.nooj4nlp.net
[21] http://nlp.stanford.edu/software/index.shtml

Figure 2.3: Stanford NLP Toolkit

Chapter 3
Question Classification

3.1 Introduction

In order to answer natural language question, the question answering process analyzes the question first to generate some representation of the information required. The processing of the question is done by the 'question analysis' phase of the Question Answering System. The question analysis phase of question processing consists of different sub-processes like, question classification, keyword extraction, derivation of expected answer type, and query expansion. This chapter focuses mainly on question classification and the derivation of expected answer type of question analysis module. The query expansion process shall be discussed in chapter 4.

This chapter briefly describes the process of question classification and its significance in designing a Question Answering System. It also discusses the methods used by Question Answering System to classify the questions and to specify the entity type expected in candidate documents, passages, or sentences. Several question patterns are presented in this chapter to fully represent and describe the Arabic questions in machine processable format. A natural language tool Nooj, has been used to design and test the patterns; results are discussed at the end of the chapter.

3.2 Why Question Classification?

Question classification represents the main part of Question Answering Systems (QAS) regardless of various types of architectures. The researchers consider question classification as very important part of a QAS (Voorhees, 2001). In addition, it has been noticed that the pursuance of question classification has made an important influence on the pursuance of QAS (Ittycheriah et al., 2001; Hovy et al., 2001; Moldovan et al., 2003).

In general, there are two motivations for question classification; **predicting the answer entity,** and **developing answer pattern** as described below:

- **Predicting Answer Entity**

To find the answer of a given question, the beforehand knowledge of the type of entities such as person, location, date etc. expected to be present in the answer is crucial. This additional piece of information helps in ranking the candidate answer passages or sentences. The question classification process helps in predicting the type of entities needed to be present in the candidate by classifying the question into various question classes (Ray et al, 2010). The answers of each of these question classes must contain specific type of entities. For instance, consider the question, " متى أصدرت لأول مرة نيويورك تايمز؟" (When the New York Times was released for the first time?) has a class of type "سنة year". While retrieving the answer for this question, the QAS will assign higher ranks to the passages containing the information about year.

- **Developing Answer Pattern**

The second motivation behind the question classification task is to develop the linguistic patterns for the candidate answers. These patterns are helpful in matching in parsing and identifying the candidate answer sentences. For instance, consider the question " من هو رائد الفضاء؟" (Who is an astronaut?), the question classification process predicts this question as "Definition" question and creates the searching patterns for specifying the answer. Some of the answer patterns for this question are: "رائد الفضاء هو ...." (An astronaut is ….) or "..... يسمى رائد فضاء" (…. is called an astronaut.) which can be better than predicting the answer entity technique.

3.3 Question Classification Approaches

The researchers have proposed several methods for question classification. In general, there are two approaches for QC: rule-based and learning-based approaches (Li and Roth, 2004).

Rule based approaches uses manual approaches for matching the questions by applying hand-crafted rules (David, 1999; Prager et al., 1999). However, these approaches are agonized from the need to generate a large number of rules (Li and Roth, 2004). In addition, rule-based approaches may perform very well on a specific dataset but may face difficulties with updated or new datasets.

On the other side, **Learning-based approaches** perform the question classification task by 1) Taking out some lineaments from the questions, 2) Train a classifier, and 3) By using the trained classifier predicting the class label. Unlike the rule based approaches, the learning approaches can be aware of recent changes in the existing dataset or can learn with a new dataset.

Some research works are using combination of the both approaches to take the full advantage of the best features of the both approaches (Huang et al., 2008; Ray et al., 2010; Silva et al., 2011).

3.4 Question Type Taxonomies

The categories of question set (classes) are usually known as question ontology or question taxonomy. In different works, various question taxonomies have been offered. The most widely used question taxonomy was proposed by Li and Roth (2002) that is based on two layer taxonomy which contains six classes of coarse-grained and fifty classes of fine-grained as shown in Table 3.1.

Table 3.1: The question classes of coarse and fine

Coarse	Fine
ABBR	abbreviation, expansion
DESC	definition, description, manner, reason
ENTY	animal, body, color, creation, currency, disease, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUM	description, group, individual, title
LOC	city, country, mountain, other, state
NUM	code, count, date, distance, money, order, other, percent, percent, period, speed, temperature, size, weight

Li and Roth (2002) published a valued set of six thousands branded questions uses by the most recent taxonomy approaches (learning-based and hybrid). The dataset is divided into two disjoint sets of questions, the first one consists of 5500 questions used for training set and the second set consists of 500 questions used as separate test set. This dataset[22] is generally known as the UIUC dataset because it was first published in the University of Illinois Urbana Champaign (UIUC). Occasionally this dataset known as "TREC" dataset, also because it is commonly used in the Text Retrieval Conference.

---

[22] (http://cogcomp.cs.illinois.edu/Data/QA/QC/)

Metzler and Croft (2005) developed the taxonomy dataset of UIUC by adding two classes (*list and yes-no-explain*). As well as they created individual 250 questions dataset taken from the questions of MadSci[23] archive. This archive is a question answering framework website that users can use.

There are some more question taxonomies for QC. Most notable among them is by Hermjakob et al. (2002) who proposed a taxonomy that contains one hundred and eighty classes and represents the widest question taxonomies that proposed till now.

3.5 Proposed Question Classification in Arabic Language

Questions in Arabic language can be classified according to the question words known as *Interrogative Particles* (IP) أدوات الأستفهام. These question words help to identify the suitable answer for a given question. IP can be classified into seven classes; some of them have two or more subclasses. IP represents the Arabic words like; kem-كَم , min-مِن, ma-ما, ayn-أينَ, mata-مَتى, ay-أي, and kaif-كَيف (**Fig. 3.1**).
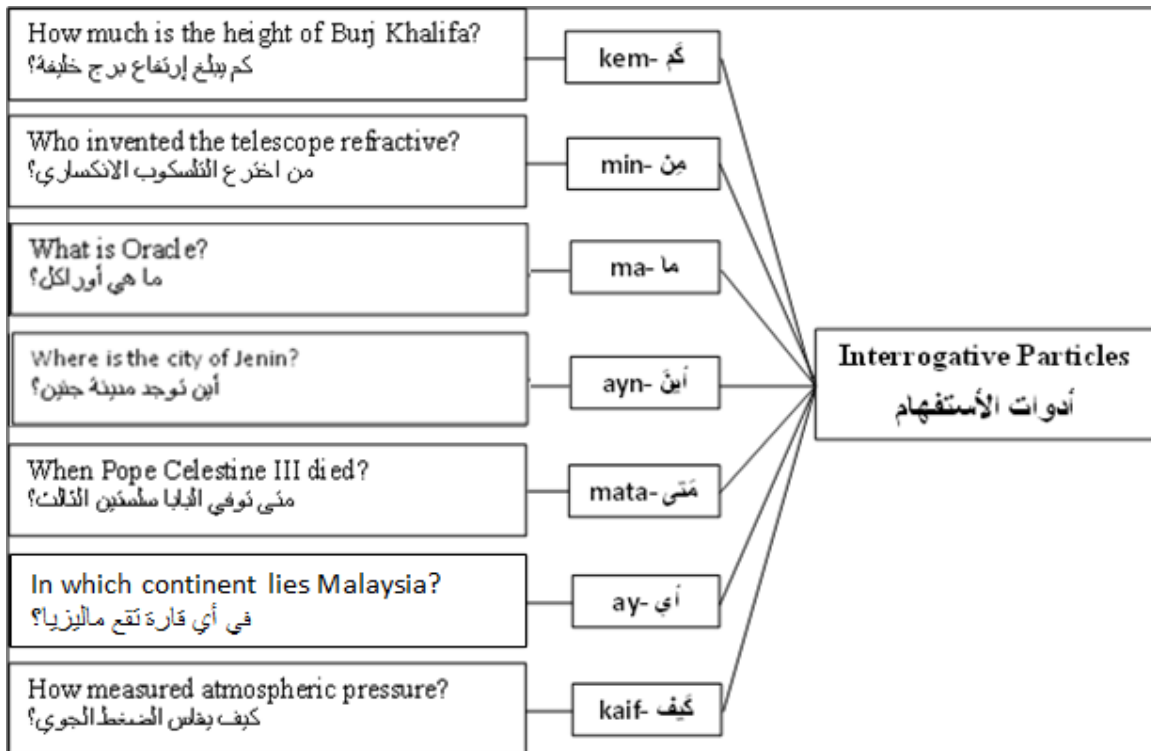

Figure 3.1: Interrogative Particles (IP) Classes

[23] (http://www.madsci.org/)

3.5.1 Logical Representation

In order to be processed by machine, the questions should be presented in a machine processable format. For the research presented in this dissertation, we have used regular expressions to represent the syntax of each IP class. Some of the IP classes may need Part-Of-Speech (POS) tagging to accurately predict the entity types expected to be present in the candidate answers. We have used Stanford Parser[24] to parse the questions. The logical representation for each IP class is illustrated as follows;

- **kem- كَم** ; It is one of the IP classes used in Arabic question answering which holds different meanings like "How much", "How many", "How long", and "How far". The classified answer type of this class is a number.

   Furthermore, this class can be classified into two subclasses; 1) depends on noun phrase (NP), for instance; كم مدينة زرتها؟ (How many cities did you visit?) and 2) the other one depends on verb phrase (VP), for instance; كم كتابا قرأت اليوم؟ (How many books did you read today?), both subclasses comes after "kem- كَم " the IP word (IPW). In all the subsequent patterns, WF (Word Form) represents set of words which do not affect the classification of the questions.

   The logical representation for all of the questions in this class can take one of the following patterns

   Pattern I:

   IPW NP VP WF

   Pattern II:

   IPW VP WF

   An NLP tool, Nooj[25], has been used to graphically represent the questions and logical patterns described in this dissertation. A brief description of these tools has been provided in section 3.6 (Nooj). The pattern used for the logical expression grammar of "kem- كَم" subclasses are shown in Fig. 3.2 for Pattern I, and (**Fig.** 3.3) for Pattern II.
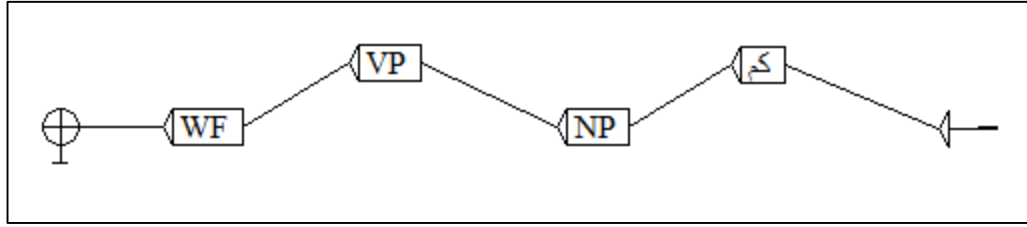

Figure 3.2: "kem- كَم" logical expression grammar (**Pattern I**)
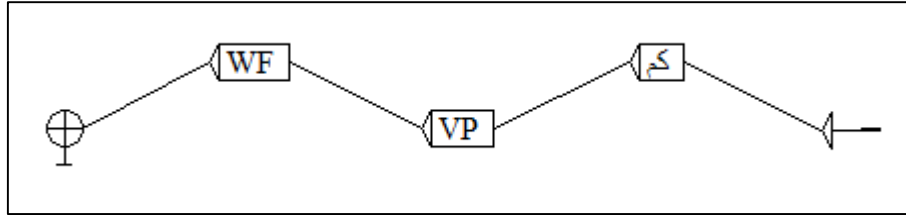

Figure 3.3: "kem- كَم" logical expression grammar (**Pattern II**)

- **min-مِن** ; It is a second class of IP classes used in Arabic question answering which is equivalent to English question word "Who" that indicated the persons, organizations , etc. For instance, consider the question; من هو رئيس الهند؟ (Who is the president of India?). The classified answer type of this class is a person. The syntax of using "min-مِن" in Arabic language is based on noun phrase that always comes after "min-مِن".

The logical representation for all of the questions in this class can take the following pattern;

IPW NP WF

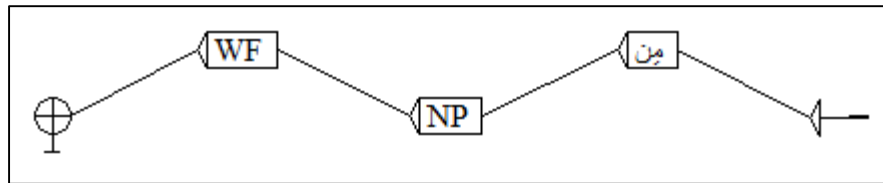The pattern used for the logical expression grammar of "min-مِن" class shown in (**Fig. 3.4**).

Figure 3.4: "min-مِن" logical expression grammar

- **ma-ما**; It is a third class of the IP classes uses in Arabic question answering that gives a meaning of "What" that indicated things, for instance; " ما هي وحدة قياس

29

"الزلازل؟" (What is the unit of earthquake measurement?). However, this class requires engaging technology definition, explanation, and clarification. The answer type of this class can be classified into different types like device, geographical location, sports, organization, art, person, etc. Furthermore, **"ma-ما"** can be classified into three subclasses; 1) depends on NP that usually starts with the word "hoa-هو", for instance; "ما هو السي ++؟" (What is C++?), 2) depends on NP that usually starts with the word "hea-هي", for instance; "ما هي اليونسكو؟ " (What is UNESCO?), and 3) depends on NP that usually starts with verb, for instance; " ما معنى بي أو سي؟" (What is the meaning of BOC?).

The logical representation for all of the questions in this class can take one of the following patterns;

Pattern I:

 IPW HOA NP WF

Pattern II:

 IPW HEA NP WF

Pattern II:

 IPW NP WF

The pattern used for the logical expression grammar of "ma-ما" subclasses shown in (**Fig. 3.5**) for Pattern I, (**Fig.3.6**) for Pattern II, and (**Fig. 3.7**) for Pattern III.
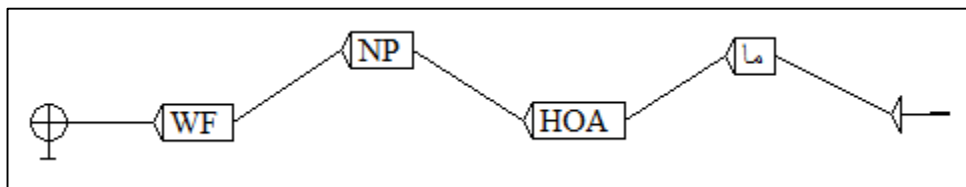

Figure 3.5: "ma-ما" logical expression grammar (**Pattern I**)
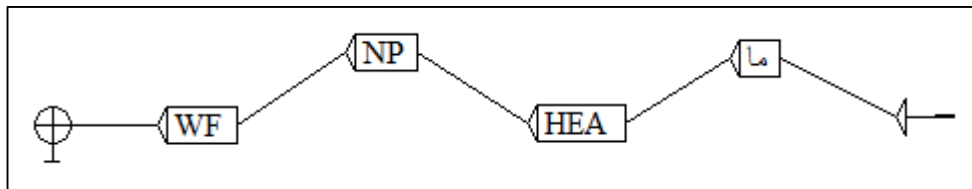

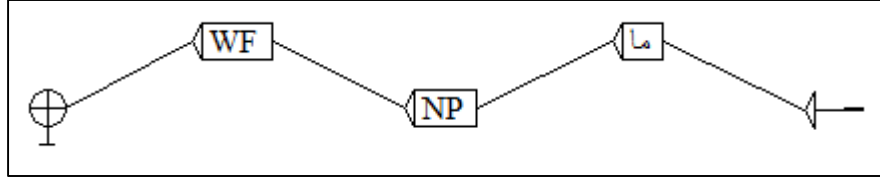**Figure 3.6**: "ma-ما" logical expression grammar (**Pattern II**)

Figure 3.7: "ma-ما" logical expression grammar (**Pattern III**)

- **ayn-أَينَ**; It is the fourth class of IP classes used in Arabic question answering which holds a meaning of "Where" that indicates the place. For instance consider the question; أين ولد ابن بطوطة؟ (Where did Ibn-Battuta born?). The classified answer type of this class is the geographical location. The syntax of using "ayn-أَينَ" in Arabic language is based on verb phrase that always comes after "ayn-أَينَ". The logical representation for all of the questions in this class can take the following pattern;

IPW VP WF

The pattern used for the logical expression grammar of "ayn-أَينَ" class shown in (**Fig. 3.8**).
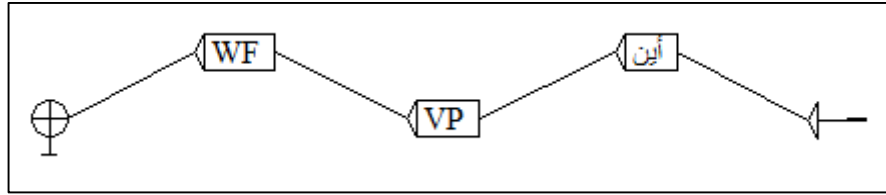

Figure 3.8: "ayn-أَينَ" logical expression grammar

- **mata-مَتى**; It is the fifth class of IP classes uses in Arabic question answering which holds a meaning of "When" that indicated the date, for instance; متى اصدرت لأول مرة نيويورك تايمز؟ (When did the first time that New York Times issued?). The syntax of using "mata-مَتى" in Arabic language is based on verb phrase that always comes after "mata-مَتى".

The logical representation for all of the questions in this class can take the following pattern;

IPW VP WF

The pattern used for the logical expression grammar of "mata-مَتى" class shown in (**Fig. 3.9**).
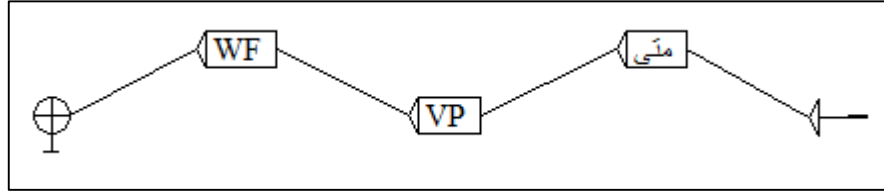
Figure 3.9: "mata-مَتى" logical expression grammar

- **ay-أي**; It is the sixth class of the IP classes uses in Arabic question answering which means "Which" that indicated sapiens and non-sapiens, for instance; " أي دولة حكمها الملك فيصل؟" (Which country ruled King Faisal?). The answer type of this class can be classified into different types like; number, geographical location, history, and sports. Furthermore, this class can be classified into five subclasses; 1) "**ay-أي**" comes after a preposition word "in-في" and followed by NP, for instance; "في أي يوم أنتخب جورج واشنطن لرئاسة الولايات المتحدة الامريكية ؟" (In which day George Washington elected to the presidency of the United States?), 2) "**ay-أي**" comes after a preposition "to-الى" and followed by NP, for instance; " الى أي مدى يصل أرتفاع برج إيفل؟" (To which extent rising up the Eiffel Tower?), 3) "**ay-أي**" comes after a preposition "from-مِن" and followed by noun phrase NP, for instance; " من أي أسم أشتق أسم مدينة قورنيا؟" (From which name the name of the city of Cyrene has been derived?), 4) "**ay-أي**" comes after a preposition "on-عَن" and followed by noun phrase NP, for instance;"عن اي دولة استقلت تيمور الشرقية؟" (On which country became independent East Timor?), and 5) finally, "**ay-أي**" comes at the beginning of the question followed by NP, for instance; " أي سنة تم بناء برج خليفة؟" (Which year was building the Burj Khalifa?).

The logical representation for all of the questions in this class can take one of the following patterns;

Pattern I: For the first five types of subclasses mentioned above;

 PP IPW NP WF

Pattern II: For the sixth type;

 IPW NP WF

The pattern used for the logical expression grammar for Pattern I shown in (**Fig. 3.10**), and for Pattern II shown in (**Fig. 3.11**).
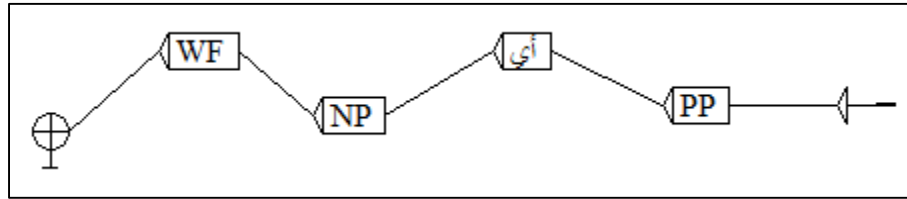
32

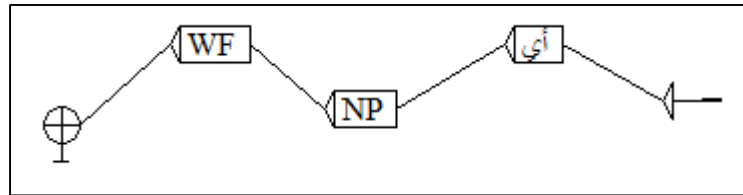**Figure 3.10**: "ay-أَي" logical expression grammar (**Pattern I**)

Figure 3.11: "ay-أَي" logical expression grammar (**Pattern II**)

- **kaif-كَيف**; It is the seventh class of IP classes uses in Arabic question answering which holds a meaning of "How" that indicated sapiens and non- sapiens, for instance; كيف يعمل جهاز الكمبيوتر؟ (How does the computer work?). The classified answer type of this class is science. The syntax of using "kaif-كَيف" in Arabic language is based on verb phrase that always comes after "kaif-كَيف".

The logical representation for all of the questions in this class can take the following pattern;

IPW VP WF

The pattern used for the logical expression grammar of "kaif-كَيف" class shown in (**Fig. 3.12**).
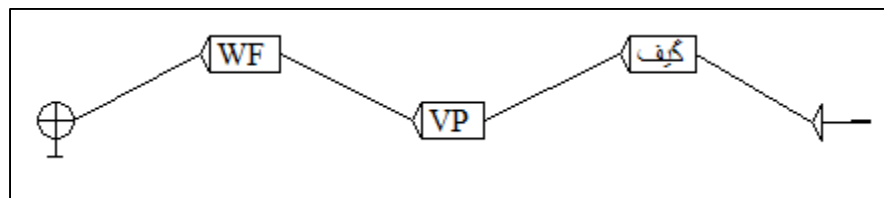

Figure 3.12: "kaif-كَيف" logical expression grammar

3.6 Nooj

Nooj is an environment that can be used to build an official images of natural language processing by applying a large corpora in real time.

The natural language can be described by formalize sets of graphs which represents the grammars of an electronic dictionaries[26]. *Nooj* helps to open dataset questions in three different languages; English, Arabic, and French.

Nooj provides tools that describes derivational and inflectional morphology, of spelling, and terminological variations. As well as, *Nooj* is used as a processing system of corpus, as it is capable to process thousands of text files (Silberztein, 2003).

In this dissertation, we used *Nooj* for processing the corpus in Arabic language developed by Y.Benajiba[27]. We used the Nooj tool to perform the following processes:

- Convert the dataset to *Nooj* text file.
- Write regular expressions using *Nooj* Regular Expression tools. These expressions include the syntax of all possible IP classes for an Arabic language. Performance Results (3.7) in this chapter shows the accuracy of using these tools.
- We applied another tools of *Nooj* for classifying the dataset questions and design a pattern that represents the syntax of each IP class. A *Nooj* Grammar tools is a graph based rules that we used for this purpose. Results show the percentage of matches' questions, more details in section 3.7 of this chapter.

3.7 Performance Evaluation of Question Classifier

Generally, the performance of a question classifier can be measured by computing the accuracy of a particular classifier on a test set. This can be defined as follows:

$$Accuracy = \frac{\text{no. of Correctly Classified Samples}}{\text{Total no. of Tested Samples}}$$

Besides accuracy, there are two other metrics which can be used to measure the class specific performance for the problem of question classification. These are; *precision* and

---

[26] (www.nooj4nlp.net)

[27] (http://users.dsic.upv.es/~ybenajiba/)

recall. In both metrics we consider *c* as a particular class (Li and Roth, 2002, 2004). The definitions of both metrics are as follows:

$$Precision[c] = \frac{no.\,of\ Samples\ Correctly\ Classified\ as\ c}{no.\,of\ Samples\ Classified\ as\ c}$$

$$Recall[c] = \frac{no.\,of\ Samples\ Correctly\ Classified\ as\ c}{Total\ no.\,of\ Samples\ in\ Class\ c}$$

As discussed earlier, we wrote several regular expressions to specify the question patterns. We used *Regular Expression* module of *Nooj* tool to write regular expression. We first tested these regular expressions over a set of 200 Arabic questions called *training questions* developed by Y.Benajiba[28] and did the necessary modification in the regular expressions. In the second phase, we wrote context free grammar for these regular expressions using *Grammar* module of *Nooj*. The question classification module was then evaluated over another set of 200 Arabic questions called *test questions* developed for TREC and CLEF.

186 questions out of 200 questions satisfied the context free grammars written by us. All of these questions were classified correctly.

$$\text{Accuracy} = \frac{186}{200} = 0.93$$

$$\text{Precision} = \frac{186}{186} = 1$$

$$\text{Recall} = \frac{186}{200} = 0.93$$

This result is promising as compared to some recent English Question Answering Systems with recall 0.63 and precision 0.7 (Samei et al, 2014) and recall 0.73 and precision 0.73 (Unger et al, 2014). Thus, our results shows the effectiveness of regular expressions and context free grammars in classifying the questions.

---

[28] http://users.dsic.upv.es/~ybenajiba/

Chapter Four

Query Expansion

4.1 Introduction

Millions of users are regularly searching over Internet using a set of keywords (called queries) to satisfy their information needs. Regrettably, these queries most probably get failed because of word mismatch between the user query and document lexicon. One of the ways to solve this problem is to expand the query by using a thesaurus which shows the semantic connection (relationship) between terms of the query (Khafajeh et al, 2013). Query Expansion, also known as the addition of the related terms, has demonstrated its effectiveness in improving the results of information retrieval process (Khafajeh et al, 2013).

Section 4.2 of this chapter describes how query expansion helps in improving the information retrieval results. Section 4.3 compiles various approaches described in literature for query expansion. Section 4 presents a query expansion algorithm to expand Arabic questions. In this algorithm, we have used Arabic Word Net (AWN) browser as an ontological resource. Ontologies are tools that able to control the knowledge beyond the concept which can be used in information search, retrieval and automatic translation (Abderrahim et al, 2013), (Al-Zoghby and Shaalan, 2015). At the end of the chapter we have discussed the results of the query expansion algorithm which was tested over a set of 50 standard Arabic questions.

4.2 Why Query Expansion?

As discussed in Chapter 1, a user enters a query into an information retrieval system and expects answers retrieved from relevant documents. The information retrieval system, in turn, identifies some of the key concepts present in the user query, and then adds variants for the key concepts which permit the information retrieval system to look for the documents that contain relevant information. This procedure faces two difficulties: first, the user usually provides the system a small number of keywords which are inadequate to distinguish between relevant and non-relevant information (Kakde, 2012). The second difficulty is the gap between the lexicon of the content creator and that of the users. The

authors of the documents may use a different lexicon to create documents on the web where users usually try to search for terms different than those used by authors which leads to failure in matching the retrieval. Furthermore, there is no clear mechanism in the traditional information retrieval system that specifies the user requirements while using the search query. For example; if the user enters a question "من قتل جون؟" (Who killed John?), the traditional retrieval system will return information about who killed John Kennedy the president of United States and information about who killed John Lennon, as well as information about other famous people with name "John" (Kakde, 2012).

From the above discussion, it is clear that one or two terms are not enough for search engines to retrieve accurate and relevant information. This creates the need for query expansion. Query expansion can add semantically equivalent terms to the original and thus enhancing the possibility of adding more documents containing relevant information. Modern information retrieval system include query expansion as necessary module to reduce the gap between the semantic and syntax of the question (Kakde, 2012).

4.3 Survey of Query Expansion Techniques

The research literature provides a large number of proposals for query expansion. All of these proposals for query expansion can be classified into three different categories:

1) *Manual*, which is mostly connected with Boolean Online Searching. Manual query expansion is performed by selecting the terms of the query for expansion manually and interpreting the topic of the query using thesaurus such as WordNet synsets (Kotov et al, 2012).

2) *Automatic*, the information retrieval system is responsible of increasing the initial or succeeding queries depending on certain methodology (uses numerous approaches classified into two main faces; *Probabilistic* and *Ontology*, both will be detailed in section 4.3.2 of this chapter) (Carpineto, et al, 2012), (Shaalan et al 2012) (Ray et al, 2009).

3) *Interactive*, both user and the information retrieval system are responsible for specifying and choosing terms required by the query expansion, this can be done by two steps; first the retrieval system use to choose, retrieve and then rank the terms of an

expansion. Secondly, the user should decide which helpful terms are required for the query from the terms ranked list (Carpineto, et al, 2012).

The expansion terms can be selected upon the input corpus or may be selected according to the external input corpus source like *ontology* or *thesaurus* (Carpineto, et al, 2012).

4.3.1 Probabilistic Methods

Probabilistic Query Expansion (PQE) usually depends on calculating the number of terms occurrence in the documents and choose the most likely terms related to the query. PQE can be categorized into two main classes; *global* and *local* methods (Manning et al, 2008). Global methods are techniques use to apply corpus-wide statistics to produce a list of nominee terms, which will be used to expand the query most alike to the query terms. The analysis of the global techniques shows that it is solid but it includes high heavy resource according to the calculations of the terms' similarity which usually use to be implemented off line. One of the primary fruitful analysis techniques is the clustering (Fernández, 2011) which grouping the document terms into clusters according to the suggested hypothesis. Queries are expanded by using this hypothesis which clusters the document terms depending on their number of occurrence in the same cluster.

On the other hand, Local methods techniques known as "relevance feedback" (Saneifar et al, 2014) refer to the process of interaction which assists to develop the retrieval performance. That means, the Information Retrieval System (IRS) returns the prior set of documents' results after the user query submission, then IRS would ask the user to judge the relevant documents. Continually, the query would reformulate by IRS according to the user's decisions and returns set of new results (Saneifar et al, 2014). These techniques make Local methods faster than Global one (Rahman et al, 2011). There are normally three types of relevance feedback; 1) explicit, 2) pseudo, and 3) implicit. In case no relevance decision found, the relevance feedback pseudo may be implemented by taking a few number of results (top ranked documents) appearing at the prior retrieval and assuming them as relevant, then enforces relevance feedback. In parallel, between pseudo relevance feedback and relevance feedback we can find implicit feedback, in which the

user's information requirement can be deduced by interacting with the system (Saneifar et al, 2014).

4.3.2 Query Expansion using Semantic Web

The relative usefulness of information retrieval systems is mainly affected by a mistake with a query consisting of a few keywords needed for the real user information. One of the well-known ways to get the better of this restriction is automatic query expansion (AQE) (Carpineto and Romano, 2012), where original query of the user is increased by new features with a parallel meaning. AQE has an extensive history in the information retrieval society but recently become a standard of scientific and experimental ripeness, exclusively in the settings of laboratory like TREC (Carpineto and Romano, 2012).

Ontology browsing is a specific language well-known AQE technique (Carpineto and Romano, 2012). Knowledge prototypes such as ontologies and thesauri deliver an income for rephrasing in context the user's query (Carpineto and Romano, 2012). In our proposed system, we used Arabic Word Net (AWN) tool for query expansion using Ontology knowledge base "SUMO".

On the other hand, (Li et al, 2007) suggested query expansion depends on Wikipedia, this can be done by using the category plans of its articles. The query works according to the Wikipedia gathering and each category is allocated a weight relative to the number of outranked articles allocated to it. Then articles re-ranked documents depending on the accumulation of weights' categories to each belonging.

Moreover, (Milne et al, 2007) proposed thesaurus-based on query expansion using Wikipedia. The thesaurus is derivative with criteria, so that the relevant topics to the document gathering are included. They proposed to get results of important topics from a query by examining successive arrangements of words in the query in contrast to the thesaurus.

4.4 Proposed Query Expansion Methodology

As described in the previous section, there are two main approaches for query expansion: Manual and Automatic. In this section we are proposing a manual query expansion approach for Arabic Question Answering Systems. The proposed query expansion

algorithm uses an ontological resource to find the semantically equivalent words. The detail of the algorithm is as follow:

Input: A user query (Q)

Output: A semantically enhanced query (Q_E)

Steps 1: Extract the keywords $C_1, C_2, …., C_m$ from the user query Q.

Step 2: For i= 1 to m

Use Ontological resource to extract top n semantically equivalent terms for Keyword under consideration. For Keyword Ci, semantically equivalent words are $C_{i1}, C_{i2}, …, C_{in}$.

Step 3: Construct a new Query using Boolean operators "AND" and "OR" as

(C_{11} OR C_{12} OR… OR C_{1n}) AND (C_{21} OR C_{22} OR… OR C_{2n}) AND …. AND (C_{m1} OR C_{m2} OR… OR C_{mn})

Step 4: End

Keywords are extracted from the user query (Q), and then the Ontology resource is looked for the top ten semantically equivalent terms for each of the keywords. Then Boolean operators "AND" and "OR" are applied to construct a new semantically equivalent search query.

To test the proposed algorithm, we selected 50 Arabic questions from a standard set of questions and answers, known as TREC & Clef Arabic questions, developed by Y.Benajiba[29]. We tested the selected questions by using Google search engine. The results of each question have been taken according to the top ten ranked results. We compared each rank result with the answer mentioned in our selected database. A comparison result of each rank has been recorded in the result section of this chapter (section 4.5 Results).

In the second phase of testing, by using the same set of questions; a query expansion has been applied by taking each keyword of the question and find its synonyms using Arabic WordNet (AWN) tool. In addition, the synonym of each word have been formalized in the question by using "OR" logical operator, then the resulting query string has been tested using Google search engine. For instance, the question;

---

[29] http://users.dsic.upv.es/~ybenajiba/

"ما هو المنصب الذي شغله أرييل شارون" (What is the position that Ariel Sharon held?), this question has been expanded by using query expansion using AWN to;

"ما هو (المنصب أو الوظيفة أو المكانة أو المرتبة) الذي (تبوأه أو شغله أو عمله) أرييل شارون"

. Here "أو" indicate logical operator "OR" and "AND" operator is default concatenation operator. Then we are collecting top ten results of the Google search engine for the modified query.

4.4.1 Arabic Word Net Tool

The Arabic Word Net (AWN) tool is a separate application that can be executed on any computer includes a Java virtual machine. It is a freely available tool to provide semantically equivalent words which can be used in many information retrieval and NLP applications. To carry out the research proposed in this dissertation we used AWN browser release 2.0 Beta version, developed by Informatics NLP Team[30]. The main motive of using AWN browser is to search for concepts that can be done either by using English or Arabic languages.

This version of AWN uses different ontologies like English, Arabic, and SUMO, where each ontology type has its interface with distinct panel. Each panel can be distributed into three universal segments; an input segment, a gloss segment and a segment for the word tree beside any extra language-specific characteristics.

In our system we checked each word (verb) of the question using AWN which includes 11269 synsets and 23481 Arabic words.

4.5 Results

This section describes the results of the proposed query expansion algorithm. To analyze the impact of the proposed query expansion algorithm, we used a standard set of 50 Arabic questions and answers compiled by Y.Benajiba[31] from TREC and CLEF as dataset. These questions were first fed into Google search engine and top ten answers for each question were retrieved. These answers were analyzed in terms of numbers of

---

[30] http://globalwordnet.org/arabic-wordnet/awn-browser/

[31] http://users.dsic.upv.es/~ybenajiba/

correct answers. For instance; the question "من كان أول رئيس للولايات المتحدة الأمريكية ؟" (Who was the first President of the United States of America?), Google search engine gives six correct answers out of first ten answers. Moreover, another instance just like "ما هو العام الذي ألقيت فيه القنبلة الذرية على هيروشيما ؟" (What year the atomic bomb was dropped on Hiroshima?) shows three correct answers out of first ten answers.

The same sets of questions were then semantically enhanced using the proposed algorithm. The Arabic WordNet browser was used to find the semantically equivalent words. The set of 50 expanded queries were fed into Google to retrieve the relevant answers. These answers were also analyzed in terms of numbers of correct answers. For instance;

"من كان (أول أو الأول) (رئيس أو زعيم) للولايات المتحدة الأمريكية ؟ "

The results show ten correct answers out of top ten answers after applying the query expansion. Moreover, another instance like;

"ما هو (العام أو الحول أو السنة) الذي ألقيت فيه القنبلة الذرية على هيروشيما ؟"

The results show nine correct answers out of top ten answers. The query expansion results as shown in the figure 4.1 indicate that query expansion has positive impact on the number of correct answers retrieved by the search engine. The average of correct answers per question we received before query expansion is 4.2 while it is 6.7 after query expansion.
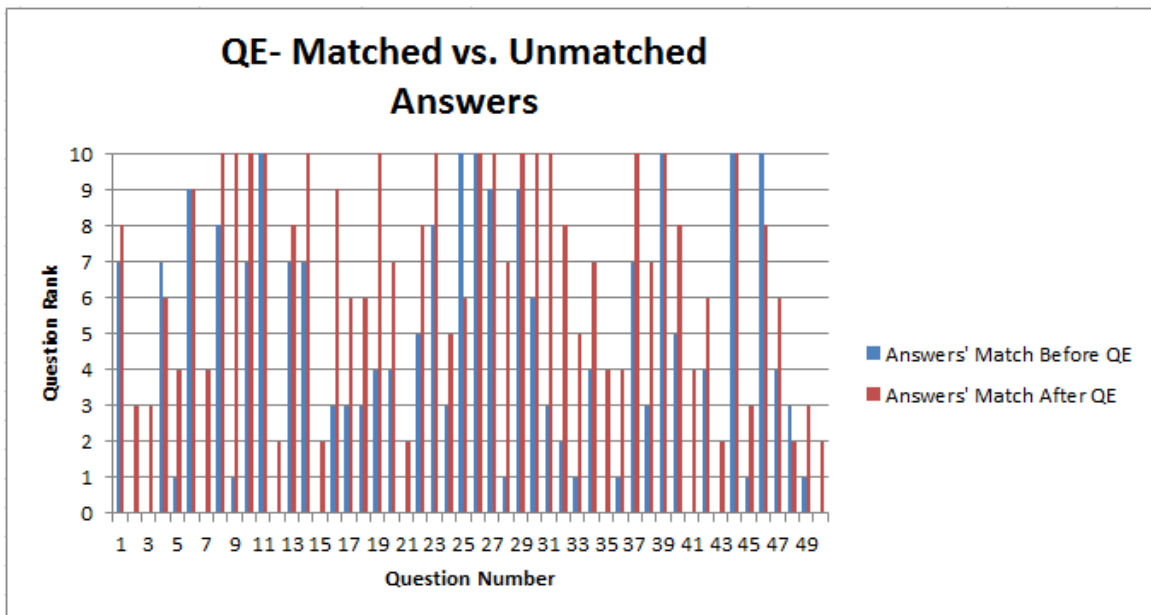
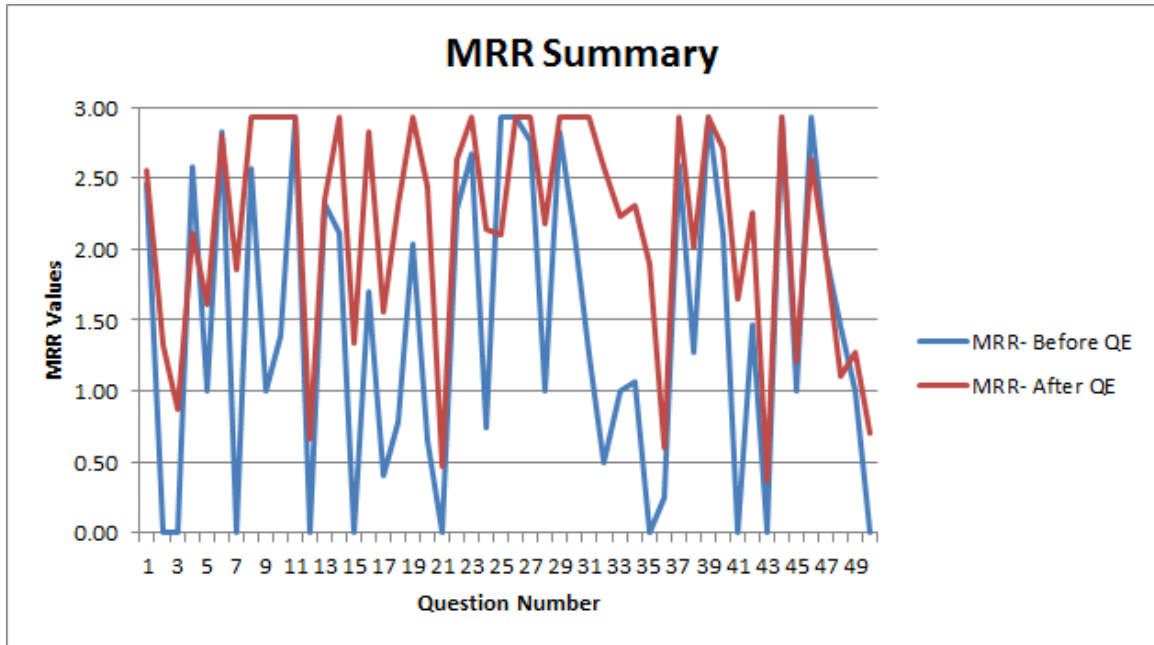**Figure 4.1: Questions summary (Matched vs. Unmatched answers)**

Mean Reciprocal Ratio (MRR) indicates how well the information retrieval systems are ranking the retrieved documents. MRR for a question 'Q' can be defined as

$$MRR\ (Q) = \sum_i 1/i$$

Where i is the rank of the correct answer. For example, if the correct answers for a question is found in documents ranked 2,4 and 8, then MRR will be ½+1/4+1/8 = 0.875. We analyzed the results of the query expansion using MRR also as shown in figure 4.2. The rank of MMR values varies from 0.0 to 3.0 for the questions under consideration in both cases, before and after applying query expansion. We can notice in general that the MRR values before query expansion fluctuated from 0 to 2.9, while some results gives good results especially questions 13 to 19 and 41 to 46.

MRR Summary towards using query expansion

The MRR average of correct answers per question we received before query expansion is 1.45 while it is 2.15 after query expansion.

Chapter Five

Conclusion and Future Work

5.1 Conclusion

Arabic Question Answering Systems could not match the pace due to some inherent difficulties with the language itself as well as upon to the lack of tools offered to support the researchers. The task of Question Answering can be divided into three phases; Question Analysis, Document Retrieval and Analysis, and Answer Analysis. Each of these phases plays crucial roles in overall performance of the Question Answering Systems. In this dissertation, the aim was to take the challenge to design and develop Question Analysis (QA) module of Arabic Question Answering Systems. The rationale behind the selection of QA was the lack of researches regarding QA for Arabic language. This phase consists of two subtasks; 1) Determining the type of questions (Question Classification) and 2) Query Expansion, where additional semantically equivalent keywords are added to the user query. In order to accomplish these tasks we used the following tools among the others;

- *Nooj*
- Arabic WordNet (AWN)

Nooj as mentioned in chapter three, is known as an environment that developed to be used to build-coverage, official images of natural languages by applying them to large corpora in real time.

Arabic WordNet is an Arabic tool similar to its English counterpart WordNet in most of the aspects and the relations. This tool has been used in chapter four in Query Expansion.

Chapter three in this dissertation described the process of question classification and its significance in designing a Question Answering System. It also discussed the methods used by Question Answering System to classify the questions and to specify the entity type expected in candidate documents, passages, or sentences. Several question patterns are presented in this chapter to fully represent and describe the Arabic questions in machine processable format. A natural language tool, Nooj, has been used to design and test the patterns and results are discussed at the end of the chapter. The research work carried out on question classification resulted in the publication in *The International*

Conference on Information and Communication Technology Research (ICTRC 2015),
(ICTRC 2015) (Al-Chalabi et al, 2015).

Chapter four in this dissertation demonstrates the meaning of Query Expansion and how to expand query expansion algorithms. Section 4.2 of this chapter describes how query expansion helps in improving the information retrieval results. Section 4.3 compiles various approaches described in literature for query expansion. Section 4 presents a query expansion algorithm to expand Arabic questions. In this algorithm, we have used Arabic Word Net (AWN) browser as an ontological resource.

At the end of each of these two chapters we have discussed the results of the proposed methods. In order to evaluate the proposed methods we have used standard datasets such as the corpus in Arabic language developed by Y.Benajiba[32], as well as a set of 50 standard Arabic questions. The performances of the proposed methods are satisfactory.

5.2 Future Work

In this dissertation, our focus was on the Question Analysis phase of the Arabic Question Answering Systems. We proposed language rules and patterns to classify questions into pre-defined question categories. This work can be extended to automatically generate patterns for question classifications using machine learning. We also proposed a query expansion method that used AWN for finding semantically equivalent words. Though AWN is useful in finding semantically equivalent words for query expansion, research can be carried out to improve content of the AWN and applicability of improved version of AWN. We focused only on designing and developing Question Analysis module of Arabic Question Answering Systems. As a future work, same can be applied for the other two phases of Question Answering Systems; that is "Document Retrieval and Analysis" and "Answer Analysis".

In Document Analysis, we can look for such methods used in information retrieval including tools, evaluation, and corpus.

In Answer Analysis, we can look for such methods used in this phase including evaluation, tools, and corpus.

---

[32] http://users.dsic.upv.es/~ybenajiba/

References

Abdelbaki, H.; Shaheen, M.; Badawy, O.: ARQA high performance arabic question answering system. In: Proceedings of Arabic Language Technology International Conference (ALTIC) (2001).

Abdel-Monem, A., Shaalan, K., Rafea, A., Baraka, H., Generating Arabic Text in Multilingual Speech-to-Speech Machine Translation Framework, Machine Translation, Springer, Netherlands, 20(4): 205-258, December 2008.

Abderrahim, M. A., Abderrahim, M. E. A., & Chikh, M. A. (2013). Using Arabic Wordnet for semantic indexation in information retrieval system. arXiv preprint arXiv:1306.2499.

Al-Chalabi, H., Ray, S., **Shaalan, K.,** Question Classification for Arabic Question Answering Systems, The International Conference on Information and Communication Technology Research (ICTRC 2015), Khalifa University, Abu Dhabi, UAE, 17-19 May, 2015.

Al-Zoghby, A., Shaalan, K., Conceptual Search for Arabic Web Content, Lecture Notes in Computer Science, Computational Linguistics and Intelligent Text Processing (CICLing), 9042: 405-416, Springer, Berlin Heidelberg, 2015.

Bekhti, S., Rehman, A., Al-Harbi, M., & Saba, T. (2011). AQUASYS: AN ARABIC QUESTION-ANSWERING SYSTEM BASED ON EXTENSIVE QUESTION ANALYSIS AND ANSWER RELEVANCE SCORING. International Journal of Academic Research, 3(4).

Benajiba, Y., Rosso, P., & Soriano, J. M. G. (2007). Adapting the JIRS passage retrieval system to the Arabic language. In Computational Linguistics and Intelligent Text Processing (pp. 530-541). Springer Berlin Heidelberg.

Benajiba, Y.; Rosso, P.: ANERsys 2.0: conquering the NER task for the Arabic language by combining the maximum entropy with PoS-tag information. In: Proc. Of Workshop on Natural Language Independent Engineering, IICAL-2007 (2007)

Benajiba, Y.; Rosso, P.; BenediRuiz, J.: ANERsys: an Arabic named entity recognition system based on maximum entropy. Com-put. Linguist. Intell. Text Process. 143-153 (2007)

Benamara, F.(2004). Cooperative question answering in restricted domain : the WEBCOOP experiment, *In Proceedings ACL 2004 Workshop on Question Answering in Restricted Domains*.

Bobrow, D., Kaplan, R., Kay, M., Norman, D., Thompson, H. & Winograd, T.(1977). GUS, a frame driven dialog system, *Artificial Intelligence*, 8, 155-173.

Brill, E., Dumais, S. & Banko, M.(2002). An Analysis of the AskMSR Question-Answering System, *In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.

Brini, W., Ellouze, M., Mesfar, S., & Belguith, L. H. (2009, September). An Arabic Question-Answering system for factoid questions. In Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on (pp. 1-7). IEEE.

Buckwalter, T.: Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, catalog number LDC2002L49. ISBN 1-58563-257-0 (2002)

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR), 44(1), 1.

Chung, H., Song, Y., Han, K., Yoon, D., Lee, J., Rim, H. & Kim, S.(2004). A Practical QA System in Restricted Domains, *In Proceedings ACL 2004Workshop on Question Answering in Restricted Domains*.

Clarke, C., Cormack, Kemkes, G., Laszlo, G. M., Lynam, T., Terra, E. & Tilke, P. (2002). Statistical selection of exact answers (MultiText experiments for TREC 2002), In Notebook of the 11th Text Retrieval Conference (TREC 2002), NIST Publication, pp.162-170.

David A. Hull. Xerox TREC-8 question answering track report. In Voorhees and Harman, 1999.

Diab, M.: Second generation AMIRA tools for Arabic processing: fast and robust tokenization, PoS tagging, and base phrase chunking. In: Proceedings of second international conference on Arabic language resources and tools, pp. 285-288 (2009)

Dumais, S., Banko, M., Brill, E., Lin, J. & Ng, A.(2002). Web Question Answering: Is More Always Better?, *Proceedings of SIGIR' 2002*, 291-298, Aug.

Ellen, M. V.(2001). Overview of the TREC 2001 question answering track, *In Proceedings of the 2001 Text REtrievalConference (TREC 2001)*.

El-taher, A., Abo Bakr, H., Zidan, I., Shaalan, K., An Arabic CCG approach for determining constituent types from Arabic Treebank, Special Issue on Arabic NLP, Journal of King Saud University - Computer and Information Sciences, Elsevier, 26(4):441 - 449, 2014.

Fernández, M., Cantador I., López V., Vallet D., Castells P., & Motta E. (2011). Semantically enhanced Information Retrieval: an ontology-based approach. Web Semantics: Science, Services and Agents on the World Wide Web, 9(4), 434-452.

Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G. & Jacquemin, C. (2001). Terminological variants for document selection and questionanswer matching, In Proceedings Association for Computational Linguistics Workshop on Open- Domain Question Answering, pp.46-53.

Giampiccolo, D., Forner, P., Herrera, J., Peñas, A., Ayache, C., Forascu, C., ... & Sutcliffe, R. (2008). Overview of the clef 2007 multilingual question answering track. In Advances in Multilingual and Multimodal Information Retrieval (pp. 200-236). Springer Berlin Heidelberg.

Green, W., Chomsky, Laugherty, C. K.(1961). BASEBALL: An automatic question answerer, *Proceedings of the Western Joint Computer Conference*, pp219-224.

Gross, J. M., Blue-Banning, M., Turnbull, H. R., & Francis, G. L. (2014). Identifying and Defining the Structures That Guide the Implementation of Participant Direction Programs and Support Program Participants: A Document Analysis. Journal of Disability Policy Studies, 1044207313514112.

Guo, Y.(2004). Chinese Question Answering with Full-Text Retrieval Re-Visited, *Master Thesis*, the University of Waterloo.

Habash, N., Rambow O., Roth R.: MADA+TOKAN: a toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In: Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, pp. 102-109 (2009)

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

Harabagiu, S., Moldovan D., Clark C., Bowden M., Williams J. & Bensley J.(2003). Answer Mining by Combining Extraction Techniques with Abductive Reasoning, *In Proceedings of the TREC-2003 Conference*, Language Computer Corporation, pp375.

Harabagiu, S., Moldovan D., Pasc, M., Surdeanu M., Mihalcea R., Girju R., Rus V., Lacatusu F., Morarescu P. & Bunescu R.(2001). Answering Complex, List and Context Questions with LCC's Question-Answering Server, The Tenth Text Retrieval Conference (TREC-10), Gaithersburg, MD.

Hermjakob, Ulf, Hovy Eduard, and Lin Chin yew. Automated question answering in webclopedia - a demonstration. In In Proceedings of ACL-02, 2002.

Hirschman, L. & Gaizauskas, R.(2001). Natural language question answering: the view from here, *Journal of Natural Language Engineering, Special Issue on Question Answering*, 7 (4), pp. 275-300.

Hirschman, L., Light M., Breck E. & Burger J.(1999). Deep Read: A reading comprehension system, *In Proceedings 37th Annual Meeting of the Association for Computational Linguistics*, pp.325-332.

Hovy, E. H., Gerber L., Hermjakob U., Junk M., & Lin C. Y. (2000, November). Question Answering in Webclopedia. In TREC (Vol. 52, pp. 53-56).

Hovy, Eduard, Gerber Laurie, Hermjakob Ulf, Lin Chin yew, and Ravichandran Deepak. Toward semantics-based answer pinpointing, 2001.

Huang, Zhiheng, Thint Marcus, and Qin Zengchang. Question classification using head words and their hypernyms. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP '08), pages 927–936, 2008.

Ittycheriah, A., Franz M., Zhu W. J. & Ratnaparkhi A. (2001). IBM's statistical question answering system. *In Proceedings 9th Text Retrieval Conference (TREC-9)*, NIST Special Publication 500-249.

Jurafsky, D., Martin J. H., Speech and Language Processing, Pearson Prentice Hall, 2009.

Kakde, Y. (2012). A Survey of Query Expansion until June 2012. Indian Institute of Technology, Bombay.

Katz, B.(1997). From sentence processing to information access on the World Wide Web, *In Natural Language Processing for the World Wide Web: Papers from the 1997 AAAI Spring Symposium*, pp77-94.

Khafajeh, H., & Yousef, N. (2013). Evaluation of Different Query Expansion Techniques by using Different Similarity Measures in Arabic Documents. International Journal of Computer Science Issues (IJCSI), 10(4).

Kintsch, W.(1998). Comprehension: A Paradigm for Cognition, *Cambridge University Press*, Cambridge.

Kotov, A., & Zhai, C. (2012, February). Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 403-412). ACM.

Lehnert, W.(1977). A conceptual theory of question answering, *In Proceedings 5th*

International Joint Conference on Artificial Intelligence, pp.158-164.

Li Xin, HUANG Xuan-Jing, and WU Li-de. Question classification using multiple classifiers. In Proceedings of the 5th Workshop on Asian Language Resources and First Symposium on Asian Language Resources Network, 2005.

Li, Xin and Roth Dan. Learning question classifiers. In Proceedings of the 19th international conference on Computational linguistics, COLING '02, pages 1–7. Association for Computational Linguistics, 2002.

Li, Xin and Roth Dan. Learning question classifiers: The role of semantic information. In In Proc. International Conference on Computational Linguistics (COLING, pages 556–562, 2004.

Light, M., Brill, E., Charniak, E., Harper, M., Riloff, E. & Voorhees, E., editors. (2000). In Proceedings Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, Seattle. Association for Computational Linguistics.

Liu, Q., & Agichtein, E. (2011). Modeling answerer behavior in collaborative question answering systems. In Advances in information retrieval (pp. 67-79). Springer Berlin Heidelberg.

Magnini, B., Negri M., Prevete R. & Tanev H.(2002). Is it the right answer? Exploiting web redundancy for answer validation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), pp425-432.

Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., De Rijke, M., ... & Sutcliffe, R. (2005). Overview of the CLEF 2004 multilingual question answering track. In Multilingual Information Access for Text, Speech and Images (pp. 371-391). Springer Berlin Heidelberg.

Mahendra, R., Larasati S. D., & Manurung R. (2008). Extending an Indonesian Semantic Analysis-based Question Answering System with Linguistic and World Knowledge Axioms. In PACLIC (pp. 262-271).

Manning, C. D., Raghavan P., and Schutze H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Metzler, Donald, and Croft W. Bruce. Analysis of statistical question classification for fact-based questions. Inf. Retr., 8:481–504, May 2005.

Milne, D. N., Witten, I. H., & Nichols, D. M. (2007, November). A knowledge-based search engine powered by wikipedia. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management(pp. 445-454). ACM.

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

Moldovan, D., Harabagiu S., Pasca M., Mihalcea R., Goodrum R., Girji R. & Rus V. (2000). LASSO: A tool for surfing the answer net, In Proceedings 8th Text Retrieval Conference (TREC-8), NIST Special Publication 500-246.

Moldovan, Dan, Pasca Marius¸ Harabagiu Sanda, and Mihai Surdeanu. Performance issues and error analysis in an open-domain question answering system. ACM Trans. Inf. Syst., 21:133–154, April 2003.

Navarro, G., Puglisi, S. J., & Sirén, J. (2014). Document retrieval on repetitive ollections. In Algorithms-ESA 2014 (pp. 725-736). Springer Berlin Heidelberg.

Navarro, G., Puglisi, S. J., & Valenzuela, D. (2015). General document retrieval in compact space. Journal of Experimental Algorithmics (JEA), 19, 2-3.

Nguyen, H. D. & Kosseim, L.(2004). Using Semantic Information to Improve the Performance of a Restricted-Domain Question-Answering System, In Proceedings of the Question-Answering workshop of TALN (Treatment Automatic de la Langue Naturally), Fes, Maroc, pp. 441-450.

Niu, Y. & Hirst, G.(2004). Analysis of Semantic Classes in Medical Text for Question Answering, In Proceedings ACL 2004 Workshop on Question Answering in Restricted Domains.

Pakray, P., Bhaskar, P., Banerjee, S., Pal, B. C., Bandyopadhyay, S., & Gelbukh, A. F. (2011). A Hybrid Question Answering System based on Information Retrieval and Answer Validation. In CLEF (Notebook Papers/Labs/Workshop).

Prager, John, Radev Dragomir, Brown Eric, and Coden Anni. The use of predictive annotation for question answering in trec8. In In NIST Special Publication 500-246:The Eighth Text REtrieval Conference (TREC 8, pages 399–411. NIST, 1999.

Rahman, M. M., Antani, S. K., & Thoma, G. R. (2011). A query expansion framework in image retrieval domain based on local and global analysis. Information processing & management, 47(5), 676-691.

Rahman, T. A. (2015). QUESTION CLASSIFICATION USING STATISTICAL APPROACH: A COMPLETE REVIEW. Journal of Theoretical and Applied Information Technology, 71(3).

Rashwan, M.A; Al-Badrashiny, M.A.S.A.A.; Attia, M.; Abdou, S.M.; Rafea. (2011). A stochastic Arabic diacritizer based on a hybrid of factorized and unfactorized textual features. IEEE Transactions on Audio Speech Lang. Process. 19(1), 166-175.

Ray, Santosh, Singh Shailendra, and Joshi B. P. (2009). "Exploring Multiple Ontologies and WordNet Framework to Expand Query for Question Answering Systems", In Proceedings of the First International Conference on Intelligent Human Computer Interaction , pp.296-305, Conference Venue: IIIT Allahabad, India.

Ray, Santosh, Singh Shailendra, and Joshi B. P. (2010). A semantic approach for question classification using wordnet and wikipedia. Pattern Recogn. Lett., 31:1935–1943.

Samei, B., Li, H., Keshtkar, F., Rus, V., & Graesser, A. C. (2014, January). Context-based speech act classification in intelligent tutoring systems. InIntelligent Tutoring Systems (pp. 236-241). Springer International Publishing.

Saneifar, H., Bonniol, S., Poncelet, P., & Roche, M. (2014). Enhancing passage retrieval in log files by query expansion based on explicit and pseudo relevance feedback. Computers in Industry, 65(6), 937-951.

Santhosh, S., & Ali, J. (2012). Discourse Based Advancement On Question Answering System. International Journal on Soft Computing.

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

Shaalan, K. (2014). A Survey of Arabic Named Entity Recognition and Classification, Computational Linguistics, 40 (2): 469-510, MIT Press, USA, 2014.

Shaalan, K., Al-Sheikh, S., & Oroumchian, F. (2012). Query Expansion based-on Similarity of Terms for Improving Arabic Information Retrieval, Eds: Shi, Z., Leake, D., Vadera, S., Intelligent Information Processing VI, IFIP Advances in Information and Communication Technology, Springer, Boston, PP 167-176.

Shaalan, K., Oudah, M., A Hybrid Approach to Arabic Named Entity Recognition, Journal of Information Science (JIS), 40(1): 67-87, SAGE Publications Ltd, UK, 2014. [Published Version]

Shaalan, K., & Raza, H.. (2009). NERA: Named Entity Recognition for Arabic, Journal of the American Society for Information Science and Technology (JASIST), John Wiley & Sons, Inc., NJ, USA, 60(8): 1652–1663.

Silberztein Max, 2003-. NooJ Manual. Available for download at: www.nooj4nlp.net

Silva, João, Coheur Luísa, Mendes Ana, and Wichert Andreas. From symbolic to subsymbolic information in question classification. Arti_cial Intelligence Review, 35(2): 137–154, February 2011.

Simmons, R. F.(1965). Answering English questions by computer: A survey, Communications Association for Computing Machinery (ACM), 8(1), 53-70. Turing, A.(1950). Computing machinery and intelligence, Mind, 59, 433-460.

Soubbotin, M. & Soubbotin, S.(2002). Use of patterns for detection of likely answer strings: A systematic approach, In Proceedings of the TREC-2002 Conference, NIST. Gaithersburg, MD.

Srihari, R. & Li,W. (2000). Information extraction supported question answering, In Proceedings 8th Text Retrieval Conference (TREC-8), NIST Special Publication 500-246.

Suzuki, J., Taira, H., Sasaki, Y. & Maeda, E.(2003). Question classification using HDAG kernel, In Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, pp.61-68.

Toba, H., Ming, Z. Y., Adriani, M., & Chua, T. S. (2014). Discovering high quality answers in community question answering archives using a hierarchy of classifiers. Information Sciences, 261, 101-115.

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

Tsur, O., Rijke, M. & Sima'an, K.(2004). Biographer: Biography questions as a restricted domain question answering task, In Proceedings ACL 2004 Workshop on Question Answering in Restricted Domains.

Unger, C., Forascu, C., Lopez, V., Ngomo, A. C. N., Cabrio, E., Cimiano, P., & Walter, S. (2014). Question Answering over Linked Data (QALD-4). In Working Notes for CLEF 2014 Conference.

Voorhees, Ellen M. Overview of the TREC 2001 question answering track. In In Proceedings of the Tenth Text REtrieval Conference (TREC), pages 42–51, 2001.

Weizenbaum, J.(1966). Eliza: A computer program for the study of natural language communication between man and machine, Commun, ACM, 9(1), 36-45. Winograd, T.(1972). Understanding Natural Language, Academic Press, New York.

Winograd, T.(1972). Understanding Natural Language, *Academic Press*, New York.

Woods, W. A.(1973). Progress in natural language understanding: An application to lunar geology, AFIPS Conference Proceedings, Vol. 42, pp.441-450.

Yadav, R., Mishra, M., Prof Sr, A., & Bhilai, S. S. C. E. T. (2013). Question Classification Using Naïve Bayes Machine Learning Approach. International Journal of Engineering and Innovative Technology (IJEIT), 2(8).

Zhang, D. & Lee, W.(2003). Question classification using support vector machines, In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.26-32.

Zhang, D.(2004). Web Based Question Answering with Aggregation Strategy, In Proceedings of The 6th Asia PacificWeb Conference (APWEB2004), Hangzhou, China, pp353-362, April.

Zheng, Z.(2002). AnswerBus Question Answering System, In Proceeding of HLT Human Language Technology Conference (HLT 2002), San Diego, CA, March 24-27.